Observations®
THE TECHNICAL JOURNAL
for SAS Software Users
www.sas.com/obs

Combinatorics with Dynamic Regression

Lloyd Lubet

Lloyd Lubet has expertise in a broad range of applications. He earned an MS in Mathematics and an MS in Industrial Engineering. He has developed complex statistical and graphical programs for prestigious Research and Development facilities such as EG&G, Los Alamos National Laboratory, and Lucent Technologies. Lloyd wrote a statistical visualization program as a certified Apple developer on the Macintosh. As a systems analyst programmer, applications included statistical quality control, inventory control, finance, and auditing. Computer languages ranged from assembler to C and from SAS to COBOL in environments supporting everything from Windows NT to UNIX V to MVS JCL and TSO. Database applications ranged from Fortran with Index Sequential Access to SQL Server, DB2, INGRES, and IDMS. About this Observations article Lloyd says, "I used almost every skill I have acquired."

Lloyd is currently researching mayan cosmology, mathematics, and mysticism with the Achi Mayan Council of Elders who live in Guatemala.

Abstract

Two simple, elegant DATA steps are used to generate all possible dynamic regression models from a constrained set of explanatory variables, lags, and an ARIMA procedure. From a complete list of all models ranked by "goodness of fit, " the best fitting models are selected. These candidate models are validated by comparison to the results of time-tested, existing models, a battery of significance tests, and judgment.

Combinatoric dynamic regression (CDR) focuses on generating all possible models (which PROC ARIMA can then evaluate). CDR also provides a method to assess the PROC ARIMA's evaluations in order to select candidate models. CDR is a purely empirical method enabling data to speak for itself. CDR provides knowledge discovery. It does not provide a method to solve any particular dynamic regression model.

This paper provides instructions on how to:

- constrain the number of possible models to a manageable size
- generate the PROC ARIMA ESTIMATE statements for each model with two DATA steps
- validate the models with the highest rank.

To use CDR, familiarity with dynamic regression and PROC ARIMA is required. Also, existing, proven models are needed to ensure that highly ranked CDR models are consistent with proven models.

Incentives for using this procedure:

- Easy to implement in 32 statements!
- Automatically establish a statistical benchmark.
- Exploration and knowledge discovery - Identify potentially predictive non-structural ARIMA models.
- Cross-validate existing structural models.
- Fine-tune the lags in complex dynamic regression models as an alternative to State Space Modeling.

Because CDR is automated, you can achieve these gains with only a modest expenditure of resources and research time. It is a brute force method. CDR is a knowledge-discovery tool and not an expert system that will magically find the "best" model.

While CDR is a powerful tool, please proceed with the greatest caution. We will discuss the need for caution here only briefly and take it up in more detail in the section labeled "Statistical Precautions."

Researchers are advised to proceed with available theory and previous results before using this approach.

Finally, this method requires cutting and pasting large amounts of text from a word processing editor into a SAS DATA step.

Definitions

Dynamic Regression: A linear regression model that often combines an ARIMA procedure with explanatory variables (or their lags). If the explanatory variables are lagged, they are often referred to as leading indicators.

Y is cointegrated with a set of explanatory variables {X} (and their lags) whenever there exists a linear regression of Y against {X} whose residuals are stationary at equilibrium. Cointegration requires only that the residuals are stationary. It does not require Y or {X} to be stationary. In a dynamic regression that relies on a moving average, the models constructed from only the explanatory variables and the AR term must be cointegrated.

SBC (the Schwartz Bayesian Criteria): $-2 \ln(L) + \ln(n)\, k$ where L is the likelihood function based on the residuals, n is the number of residuals, k is the number of free parameters. We will use the SBC as the measure of goodness of fit. " It has no statistical measure for erroneous inferences and implications. In short, it has no alpha statistic. Hence, while this is an excellent tool for ranking numerous models, it is a very poor tool for measuring the reliability of any particular model. You will need to apply other methods, statistical methods to measure the probability of any erroneous significance of a particular model.

Model Complexity: The number of estimated parameters.

Alpha: The theoretical probability that a statistical test demonstrates a good fit or a fit much better than mere guessing but in fact has given you bad advice. Most statistical tests provide an alpha default to 5%. An alpha of 5% implies that each model deemed significant by this test has a 1/20 (5%) probability of an incorrect inference.

Alternatives to CDR

PROC STATESPACE or Bayesian VARS will also attempt to find an optimal subset of explanatory variables with suitable lags and will account for interaction between the variables as well. When unable to constrain the number of explanatory variables and their lags to a reasonable size, then the researcher should consider other search methods such as genetic algorithms, genetic programming, simulated annealing, or cellular automata.

Contents

- **Introduction**
- **Statistical Precautions**
  - *Constraints - Limiting the Potential Complexity of the Models*
  - *Detecting Statistical Artifacts – Avoiding Bad Models*
  - *EXAMPLE*
  - *A Suggested Battery of Significance Tests*
- **The Building Blocks of Combinatorics**
  - *Practical Examples: Doing a Forecast*
  - *Establishing a Benchmark for Model Accuracy: The Univariate BJ ARIMA Model*
- **Running the Combinatoric Dynamic Regression**
  - *Generate the List of Combinations of Models (the mod file)*
  - *Save SAS/ETS Statements*
  - *Comparison of CDR Explanatory Model and the BJ ARIMA*
  - *Example 1: An Implementation of a Full Dynamic Regression Model with MA Terms*
  - *Example 2: Models with AR Denominator Factors*
- **Conclusions**
- **References**

Introduction

Combinatoric dynamic regression is the union of combinatorics and dynamic regression. With properly defined constraints, it will automatically generate and run all possible dynamic regression models, exploiting the incredible speed and reliability of the SAS DATA STEP and PROC ARIMA in SAS/ETS software. Reader familiarity with ARIMA concepts is assumed.

The Six Stages of Combinatoric Dynamic Regression

1. Define constraints
   - Select a meaningful set of explanatory variables
   - Define maximum lags for the explanatory variables
   - Define maximum lags for ARIMA terms
2. Specify the model format
3. Generate the list of possible models with an identification number for each model (for easy reference)
4. Evaluate the ESTIMATE statements representing the possible models with PROC ARIMA
5. Create a Ranked List whose fields are each model's identification number and the resulting SBC of the estimated model; this list is sorted by SBC.
6. Validate and diagnose interesting candidate models; compare the CDR results to univariate BJ ARIMA and existing, time-tested model results.

Statistical Precautions

Constraints - Limiting the Potential Complexity of the Models

Without adequate constraints on the number of explanatory variables and their lags and the complexity of the ARIMA component, the number of combinations of possible models will increase explosively. It may take considerable skill to provide suitable constraints. To reduce the complexity of a system, the researcher may apply principal components or the more powerful projection pursuit neural net. If the researcher is unable to constrain the complexity of the models, then it is advisable to try genetic algorithms, cellular automata, or simulated annealing.

Constraining the model complexity is normally beneficial, but for CDR, it is essential. This requirement has several advantages in forecasting: parsimonious models generally outperform complex models; simple models are easier to understand and are easier to justify; and with a simple model, there is less risk of collinearity, which may bias your coefficients and statistical tests. To measure collinearity, use the COLLIN option of the FIT statement in the MODEL procedure.

Detecting Statistical Artifacts – Avoiding Bad Models

Process (P)	A system with multiple time-series inputs X1, … Xn and one time-series output Y**.** The output Y is often called the response variable. The inputs, X1, … , Xn, are called explanatory variables. When past values of the response variables are used as inputs, they are called autoregressive / moving average terms in the model denoted Y-1, Y-2, Y-3 where –1 indicates a lag of 1 time period, –2 indicates a lag on Y of 2 time periods and so on.
System	An interacting or interdependent set of variables forming a unified whole. This paper will deal with systems with only a single response variable and will emphasize the interaction and dependence of the response variable to remaining variables, often referred to as explanatory variables. Researchers requiring a broader treatment of systems, where all of the variables can interact and effect the behavior of other variables, might consider PROC STATESPACE, possibly PROC SYSLIN, or even simultaneous, stochastic partial differential equations. While identifying the system, analysts must determine which variables to include, deeming others as either unimportant, impossible to measure, or omitted for theoretical reasons.
System Context	This set of excluded or even unknown variables determines the system context. In other words, they are variables not included in the model. Excluded variables found in the context may have a powerful but unknown impact on a system. For example, changes in the policy of a presidential regime could have a dramatic impact on a relevant sector of the economy. However, this variable may be impossible to measure and predict.
Domain (D)	Let d be vector whose components n+1 components are Y, X1, …, Xn. The domain D is the set of all naturally occurring or possible d's. Every possible sample is a subset of D. The phrase "naturally occurring" implies that each variable has a normally occurring range, an expected maximum and minimum under suitable or normal conditions. The vector space formed by D will probably be in the shape of a hypercube whose limits on each side are set by the range of each component variable.

Non-Random Process (NP)	A system whose output produces predictable results given certain inputs. A system whose uniform behavior over its Domain D can be portrayed by a smooth curve algebraically represented by an ARIMA-X equation.
Random Process (RP)	A process whose output is a random series of numbers regardless of the inputs. A random process may produce samples that temporarily appear to contain a pattern. Statistically significant models can be based on such anomalous samples. These models, lacking any predictive value, are sampling artifacts. They are the results of bad data, bad sampling design, or just bad luck.
Models	Linear equations representing an underlying, repeatable process that generated the given sample. Thus, an adequate model must explain and predict most possible samples in its domain. In other words, we are trying to generalize from predictable behavior in a given sample to a model that fits an underlying, smooth pattern found in the sample and then use this derived model to fit (in some sense predict) the data found in other relevant samples.
Alpha Error - Type One (Alpha-1)	Given a sample S, run a Random Walk test on the sample output Y to make certain there are patterns worth modeling. We declare a sample to be non-random if there is less than a 5% chance that a random process could produce the non-random patterns found in sample S. Alpha is the probability that a statistical test purporting to demonstrate non-random behavior is in fact wrong. Many artifacts exist in data sets mistakenly categorized as non-random. In most research situations, there is a 1/20 chance that a statistically significant test is in truth the result of a random process and hence an artifact.

If a multivariate process P is non-random, then the researchers try to model the process.
The model assumes: if the pattern occurred in the existing sample, it will extend across most other relevant samples. This lends the model predictive power. In other words, the model is derived over the local behavior of one sample and applied globally to all other possible samples. We hope a model that fits a given sample extremely well is more likely to fit other samples as well.

Statistical Artifacts

An artifact is a model that coincidentally fits only the data found in sample S but does not represent the underlying, non-random process that generated S. An artifact often provides excellent fits only on the given sample data but fails on all other samples generated in future activity. Also, a battery of significance tests may all give misleading results. Models whose assumed reliability rests upon misleading significance tests are artifacts. In other words, by coincidence, a spurious model fits the sample data so well that the analyst is mislead into believing the model represents the underlying process. Statistical artifacts are due to the inescapable uncertainty found in most systems. We must accept some level of alpha that a given pattern in a sample is not a phantom of a random process. Other causes for statistical artifacts can be model over-specification, collinear explanatory variables, regime shifts, outliers and interventions, categorical data, highly clustered around specific behavior, unstable ARIMA terms, insufficient sample size, and the laws of chance.

Apparently, there is no escaping uncertainty

For example, weather is a chaotic system. Even a hard laboratory science such as quantum physics has the “uncertainty principle”. The great physicist Nils Bohr concluded that the ultimate, underlying reality of nature is

unknown and unknowable. In mathematics, the only deductive science, uncertainty abounds in Zorn's Lemma, Godel's Theorems, and Cantor's work on infinite sets.

In real-world applications, researchers usually do not have the luxury of scientific laws. Without prior experience or easily repeated and strictly controlled laboratory testing, an artifact may easily go undetected.

Early letters by Bayes explore the philosophy of science. He asks, "Are poor forecasts produced by bad models or merely poor measurements? If rules are made to be broken, then modeling assumptions are certain to be ignored. Or worse, the assumptions by which regression was derived are mathematical conveniences driven only for algebraic simplification and not based on reality. "

Einstein complained, "Deterministic processes may appear to be random because of missing variables." In practice, it may be impossible to account for all relevant variables.

Even in controlled experiments, lack of time or funding may lead to insufficient sample size, errors in measurement, or data entry.

However, in the face of insufficient data and the risk this implies, the analysts should rank models by their statistical significance. After all a model with strong statistical significance has a better chance of being reliable. Hopefully, the sample data is truly indicative of the average sample in the domain and likewise the model represents a pattern found in the test sample that extends to most other relevant samples. In other words, the model fits the behavior of the process almost everywhere.

The better a model represents the past behavior of a system, the more likely it will produce better forecasts than models that do not fit the existing data. Major assumption: context of the system has not changed radically from the past to the future. We are relying on uniformity of behavior of the system over its domain. This problem of uniformity over its domain plagues all models: the behavior of dynamic systems change frequently. Thus, old models and the new ones are all at risk. We endeavor, therefore, to select the strongest models tested upon samples including the most recently collected data.

Whether you select one candidate model from a short list of hand-picked candidates or you select one model from a complete list of models generated by CDR, the problem of statistical significance and artifacts remains the same. You don't know if current sample data is indicative of the underlying process driving the market. You don't know how the system and its processes will change in the future. Old models deemed reliable may suddenly fail while new models viewed with distrust may later produce effective forecasts. Uncertainty prevails. No one knows enough to predict the future. In addition, ARIMA-X, our model type, may not be appropriate as a forecasting method for a process. With only one sample, only one time series, it may be impossible to determine whether a process is totally random, chaotic, or nonlinear. Thus, there is also uncertainty concerning the adequacy of the modeling methods used.

This uncertainty applies to every candidate model regardless of the length of the list from which it was selected. Statistical significance of model adequacy is based strictly on sample data and the particular candidate model you have selected to represent patterns revealed by that data. Moreover, that statistical significance test has a 5% chance of being wrong. All candidate models must be based on the same sample data and must be fairly assessed by the same tests for adequacy and the measures for alpha. These test results are independent of all other models under consideration. Models must be tested separately on their own merits. Any significant test on any model has that same 5% chance of being wrong.

With CDR, we are flying strictly by instruments, without a controlled laboratory experiment or scientific theory. Similarly, if your airplane is out of fuel and there is a dense fog surrounding you, then by instruments you must land.

If you are a small private investor trading options in 15 minute intervals, then by statistical instruments you must decide.

The uncertainty is present regardless of the size of the list of possible candidates. As noted, even models based on theory contain risks. How do analysts with a theory reconcile a sample that theoretical models fail to fit or explain? Most theories depend on assumptions. Does the sample and the model completely satisfy the conditions?

CDR can dramatically improve our chances of finding reliable models

Does the increased number of models generated by CDR assure better reliability?
CDR takes into account all possible models. CDR ranks the models by goodness-of-fit and measure of statistical adequacy. Because CDR produces a ranked list of all possible models and ranks them by SBC, you need to consider only the top performers, say the best 10 or 20. Because, you are considering all possible models, you are very likely to find new and more adequate models. Even with a theory in hand, you can find better models that also satisfy the theory. Because the long list produced by CDR contains all possible shorter lists of models generated by non-CDR procedures, the strongest candidate models generated by CDR are likely to be more accurate and have stronger significance test than any model found in an arbitrary short list of models. Certainly, it will produce models at least as good as your favorite time tested model or panel of models or a short list of hand picked models because CDR's long list will contain all of these cases as subsets. You will consequently learn more about the system you are modeling, and it guarantees that you will be considering the most significant models. There will be no trade off between quantity and quality. CDR gives you both.

Artifacts viewed in the light of inherent uncertainty in statistical significance testing

If a significance test produces a positive result, there is an acceptable risk of 1/20 chance that the test is wrong. Hence, selection by a significance test exposes us to the risk of selecting seemingly reliable models, phantoms of chance.

It has been argued by certain fellow economists that more models means more artifacts. True.

With most complex, dynamic systems, theory is not strong enough to specify a "correct" model or even determine if a highly ranked model may be correct. That is, most applications do not have the luxury of physical laws. This is especially so in financial and econometric models. Thus, we must rely on statistics, what the data can tell us, despite the flaw of acceptable risk of misleading test results.

Also, CDR relies entirely on empirical results; that is, the data and only the data speaks for itself. You can always eliminate highly significant CDR models because they do not conform to current economic or other scientific theory.

EXAMPLE

CDR can easily generate 20,000 candidate models. With a 1/20 chance of false positives (ALPHA=.05), there may be 1,000 statistical artifacts scattered throughout the complete list of generated models. Because you will be selecting only the top fitting models, there is a better chance that these will not appear in the 10 "most significant" models. Artifacts usually do not adequately fit the sample data and possess strong significance tests. However, there can be no guarantee.

The ranking method used in this paper, SBC, has no significance test, unfortunately. There is no measure of the probability that a top performing candidate is actually an artifact. For this reason, use a battery of significance tests to gauge the probabilities of spurious, highly ranked models.

For example, let y be a dependent variable and x1–x20 be possible independent variables. Constrain the models to a linear equation of one variable and a constant. The list of possible models is

*Y=A1*X1+B1;*
*Y=A2*X2+B2;*
*Y=A3*X3+B3:*
…
*Y=A20*X20+B20;*

The model selection criteria could be as follows: rank the models by t score on Ai (where i runs from 1 to 10) and select the two models with the highest t scores, models 1 and 2. Suppose both t-tests suggest an alpha less than .05 (there is only a 1/20 chance of an incorrect conclusion).

For convenience, suppose the models with the highest t-scores are
*Y=A1*X1+B1* and *Y=A2*X2+B2.*

A t-test with an alpha of .05 means that the process generating the data was actually *Y = B1*, that is A1 is actually equal to 0. Because *A1 = 0* has an empirical alpha of only 5%, this is an artifact. The odds of the t-test being wrong is 5%. The chances of B1 equalling 0 is 5%. The test implies that B1 is most likely not zero. There are a lot of plausible, non-zero coefficients for the linear model. Linear regression merely solves for the most likely coefficient fitting the data. A significant t-test does not say that A1 is the exactly right coefficient. Instead, it says, the coefficient A1 is more likely to predict the behavior of Y based on the behavior of X1 plus a constant (referred to above as B1). There is a 5% chance the t-test is wrong. That is, the underlying process generating the data is *Y=B1*, where actually A1 = 0. It does not say that there is 95% chance that A1 is exactly right. It merely implies that it is extremely unlikely for A1 to equal 0.

If the second model, *Y=A2*X2+B2*, also possesses a t-score whose alpha is .05, then the chances of A2 actually being zero is 5%. The number of candidate models with t-tests with an alpha less than or equal to .05 does not change the probabilities on any one model being an artifact. The t-test takes each model and compares it to the evidence. Its merits are considered independent of the other models under consideration. Their existence is irrelevant.

Now suppose we increase the number of possible models to 200 linear models of a single explanatory variable:

*Y=A1*X1+B1*
*Y=A2*X2+B2*
*Y=A3*X3+B3*
…
*Y=A198*X198 + B198*
*Y=A199*X199 + B199*
*Y=A200*X200 + B200;*

The list of possible models in the prior example is a subset of the list of possible models in this example. Thus, we must produce results at least as good as those found in the prior example.

Suppose we select models whose t-score on the linear coefficient must have an alpha less than or equal to .05. How many candidates will we get? At least two (from the previous example) and probably more.

Suppose, out of these 200 models, there are now four candidates. This does not imply that the possibility of having an inadequate, incorrect model is double the chances of the previous example, which had only two candidate models. Each model, based on its own merits, has the same statistical significance, an alpha of .05.

Actually, a larger list of candidate models increases our chances of picking a more adequate model. Some of the candidates may actually have better fits; they may have coefficient t-tests whose alpha's are much less than .05. Some may have alphas of only .01 and are therefore more significant. Note: Highly significant top-ranked models are less likely artifacts and more likely reflect the process generating the sample data.

Also, new candidates with excellent significance tests may suggest models previously unsuspected by results found with a smaller base of models.

Therefore, doubling the size of the list of possible models does not double the chances of selecting an incorrect model. Doubling the list of possible models may, however, double the chances for statistical artifacts to appear in the list. The significance test applies to each candidate independent of all other models. We select the models that are most significant. The length of the list increases our chances of finding the candidates with the highest individual significance.

The test is applied to each model. The test is independent of all other models under consideration. The more models, the better your chances of finding a better fitting model, a model with better significance tests. If the model you select has a t-test with an alpha of .01, it doesn't matter if you selected your model from a list of 10,000 candidates or 10 candidates. The probability of your selected model having a non-zero coefficient remains .01. You judge each model upon its own merits.

A large list of possible models is more likely to contain better fitting models than a subset of this list.

As a strictly empirical method, CDR relies entirely upon patterns within the data. The resulting PROC ARIMA estimates of each model's coefficients are the ***most likely*** estimate given a sample of data and a model assumption such as linearity of the underlying system. Such estimates tend to minimize the errors between the fitted and actual values. However, "most likely" does not imply exactly correct estimates. In fact, it implies nothing about correctness of the estimates representing the underlying process generating the sample data. The most likely estimates merely outperform all other estimates by a goodness-of-fit criteria such as the Mean Squared Error or the SBC.

However, every model is a potential artifact whether we choose it from a short list of only 10 models or a long list of 20,000. Given a specific type of model, suppose 10% of the models are artifacts. Hence, in a short list of 10, there may be 1 artifact, whereas in a long list of 20,000, there may be 2,000 artifacts. On the other hand, both lists may contain no artifacts or all artifacts. There is no way to know for certain.

The long lists afford you more protection. The long list contains all short lists. Thus, all artifacts in the short list will appear in the long list. If an artifact also produces strong significance tests, you are more likely to select it from the short list. The long list is likely to produce models with goodness of fit and significance tests better than the artifact selected from the short list. Therefore, the best performing models will perform at least as well and probably substantially better then models from any short list.

We can use the significance tests to approximate the probability of an artifact in top performing models.

Given the model: $y = a * x + b$, we want to test the probability that a > 0 is not an artifact. This test is called a hypotheses test or a significance test. The process pretends the value for $a = 0$, that is, x has no real effect on y. So, we assume $a = 0$ and use a t-test to compute the probability that the odds that the coefficient "a" does not = 0 could occur by chance giving a actually = 0. If the odds of this occurrence for this data sample are low enough, we accept the alternate hypotheses, that is, a does not = 0, x has an impact on y. Note: When the probability that a does not = 0 is low enough, we reject the null hypothesis. The question is, what odds are low enough to justify accepting that "a" is probably not equal zero. The default is usually 5%. That is, given the underlying process really has an "a" = 0 but the model and its associated significance test are not zero, we accept the fact that "a" does not = 0. We say that the odds are too low for this data sample to be generated from a process whose "a" actually does = 0. We are accepting a 5% chance of being mistaken. This probability of being mistaken is called the alpha, and the alpha is the probability a candidate is an artifact.

A large number of models will generate more high performing models whose alpha is less than 1%. This means, a is not equal to 0 and the significance test implies that the odds of this data being generated from a process whose A actually = 0, is only 1%. There is only 1 chance in 100 to draw a sample form this null process that produces a significant A. This occurrence is highly unlikely so we usually accept the hypotheses that A does not equal zero.

Again, large samples and more models simply reduce the size of the alpha, the chances of an artifact.

Suppose we accept any model in our list whose alpha is less than 5%. Then out of each 20 models (deemed significant), we run the risk of accepting an artifact. If our list is short, say 20 models, then there may be one artifact found in this list. If we generate 20,000 models and accept candidate models with an alpha < 5% then there may be 1,000 artifacts, models that test well and fit well.

Hence, CDR, which produces lists of at least 20,000 alternative models, will produce more artifacts than a short list of only 20 models. If so, why use CDR?

The more models you generate, the lower the alpha is likely to be. A list of 20 may generate a model whose alpha is .05 but a list of 20,000, which will probably locate all of the good fits, may generate many alphas < .05, such as .01 or .001. Significant fits like these are extremely unlikely to be artifacts.

A longer list of candidates is likely to reduce the alpha. This is a key benefit of CDR. We see all models and select those with the best fits and the lowest alphas. Therefore, our top performing candidates are least likely to be artifacts.

The ARIMA regression models try to pick the most likely parameter estimates that minimize the total error between the model predicted values (fitted) and the actual values. The closer the total of the absolute fitted values agrees with the actual values, the better the model explains the sample data, and the more likely the model truly represents the process driving the sample data (assuming the sample is indicative of the underlying process). In most ARIMA algorithms, the absolute difference between the fitted and actual all have equal weights. If the actual data has several outliers, huge deviations from the underlying trend, the more likely that artifact models will be accepted.

First, graph the actual data. Examine them for regime shifts and huge outliers. If so, you can add indicator dummy or intervention variables to represent these severe, abrupt deviations from the underlying pattern. For example, if a huge deviation occurs at point 18, then you can add a dummy variable that is all zeros except at point 18, where it is a 1. This will alleviate the powerful biases outliers exert on model significance tests and goodness of fit. Frequently, this simple technique will force almost all artifacts from consideration as candidate models. Their significant tests and goodness of fit will become inadequate.

Measures of goodness of fit measure how closely the fitted values approximate the actual values. The sum of the differences between the fitted and actual values is called the residuals. Graph the residuals. Are the differences, the individual residuals, uniformly tight across all cases of the sample? Are there small isolated segments where the residuals are tight and vast areas where they are not? This can indicate a regime shift where the model fits under certain circumstances and fails utterly in most other circumstances. This model may be an artifact or indicate the need for dummy variables.

Replacing the R-square is the more elegant SBC. The time series could be a non-recurrent string of random numbers. The model may select too many parameters. The more parameters used, the tighter the fit, generally, and the better the significance tests. In polynomial regression, it is possible to fit even a series of random numbers by increasing the degree of the polynomial, the number of model parameters. The SBC penalizes models for complexity, that is, the number of parameters. It favors simpler models.

The SBC takes into account the maximum value for a likelihood function (measured by the residuals errors), the sample size (always an asset), and the number of parameters used in the model (always a deficit). The SBC has proven more robust than the mean square error (the simplex goodness-of-fit models) and the adjusted R-square. The SBC allows you to compare models in a long list of models with the top performing models having the lowest SBC. Generally, a model with a low SBC is much less likely to be an artifact than a model with a high SBC. SBC ranking of long lists will select the most robust and reliable models. A short list may not even include the best models.

In summary, better models have higher significance tests, lower sum of squared residuals (or higher goodness of fit), and lower SBC scores. The more models the better. A long list is more likely to contain a model representing the underlying process than a severely limited short list.

Suppose revenues are actually dependent upon and generated by interest rates and Gross National Product. Suppose you opt for a short list, leaving out GNP. Hence, this short list will never completely explain the data or its underlying generating process. CDR will include models using interest rates but not GNP. That is, its long list will subsume the possible shorter lists, lists using fewer possible variables.

The CDR generates models using all combinations of all relevant variables. No robust models are left out by accident. The CDR includes all shorter lists. If you have considerable knowledge on a process, your resulting short list will be included in the CDR. Also, the combinations of lags on explanatory variables and ARIMA structure can be overwhelming. CDR will consider all possible lag structures automatically. Again, you consider all models while eliminating the possibility of missing many valid models. Because the CDR lists all possible models, the models with the best goodness of fit and the fewest estimated parameters will be at the top. A short list may miss some of these options and be less robust or valid.

The longer CDR list may also point out strong models worthy of further study. Most models have thousands of combinations of explanatory variables and lags.

Probably, you are forecasting the same process frequently. Compare your new models to your tried-and-true models. Do your new candidate's forecasts agree with the time-tested models? If the new models do not concur, then keep them separate and test them on paper for an extended period of time. Make the new model prove itself.

Final note: Whether you use only one model or 10,000 models, the risks are exactly the same. CDR insures you are not missing out on models with greater reliability and predictive power.

Bad Data Can Produce Bad Artifacts

Insufficient sample size often leads to statistical artifacts. A small sample may not be indicative of the normal behavior of a process. A single sample out of millions of possible samples may be a fluke, a rogue if you will. Bad samples produce bad artifacts.

Biased sampling strategies will also lead to abnormal samples by design. Hence, any model fitting this sample may not fit the data in most other (normal) samples.

Cluster sampling with a poor sampling strategy may lead to very dense data structures with highly localized information, which in turn often lead to poor generalizations about the process' global behavior. For highly clustered data, the resulting model usually does not extrapolate beyond the sample.

Outliers are anomalous points of data. They can arise from data entry error, bad news, or surprising economic statistics. Because the regression tries to minimize the square of residuals, outliers, which stray far from the underlying pattern, will have the highest residuals. The algorithm will place more weight on residuals associated with outliers. The model may fit the outliers better than the remaining data. Outliers are likely causes of artifacts. A good model best represents the normal behavior of the system. Outliers may bias that focus toward the abnormal behavior often due to exogenous factors.

We assume a system behaves uniformly over time. It must be consistent and have a smooth, underlying trend. By the Central Limit Theorem, most samples will cluster around the mean system behavior and these samples most likely produce models that best represent the system's average behavior. When a system's structure and behavior change abruptly, this change is called a regime shift or an intervention. Without taking precautions, our single model may attempt to fit two different patterns of behavior. In most cases, the model will produce a weighted average of the separate models representing each regime. The weights are determined by the percentage of the sample found in each regime. Such compromised models are usually inadequate and also produce poor forecasts.

This section discusses detection or avoidance of artifacts. Repeated sampling is the best protection. If we cannot acquire new samples and the original sample is deemed large enough, then we can run the model against sub-samples. These schemes are called jack-knifes and bootstraps. Their use is controversial and they will not be discussed further.

Artifacts can also fail to accurately model subsets of a given sample.

Leave out the last four points and use the model on the remaining data to predict the known but missing points. You could also repeat this process by reversing the order of the time series and rerunning this test. Purportedly, adequate models should be able to closely approximate or predict the missing values in either direction because we must assume the underlying process has not changed over time.

SAS/ETS PROC ARIMA provides an additional, powerful tool. Researchers can also use the back-forecasting option BACK. For example, if you specify a LEAD option equal to four time periods to be forecast, you could instruct PROC ARIMA to start forecasting before the end of the input data. This simulates forecasting data, allowing you to measure the model's performance against known outcomes. Compare forecasting results with the equal BACK and LEAD values of existing models and candidate models.

A Suggested Battery of Significance Tests

To evaluate a candidate model from the CDR list, test the residuals for model adequacy. That is, test how well the model accounts for the behavior of the dependent variable, with the Ljung-Box Q test, the Random-Walk RW test, the Generalized Durbin-Watson test. Plot the AFC, IAFC, and PAFC. They should all decay rapidly to values statistically equivalent to zero, that is, autocorrelations of the residuals exceed +/- 2*sqrt(T), where T is the sample size.

The Ljung-Box, one of the most widely used significance tests for model adequacy, examines the residuals over time. It tests for departures of the error's ACF (auto-correlation function).

AutoReg provides the RW and higher order Durbin-Watson diagnostics.

The SBC balances goodness of fit against simplicity. More complex models with more estimated coefficients are penalized more heavily than simpler models. However, parsimony is a principle for model selection, not a physical law. The underlying system may require a complex representation. Hence, check candidates ranked by SBC against the same list of models ranked by the Mean Squared Error (which is purely a measure of goodness of fit without regard to parsimony).

Check the R-squared, the percentage of the variation in the dependent variable explained by the model against expected or required forecast accuracy.

Compare the coefficients and lags of each input variable in the new candidate model against corresponding coefficients and lags in existing models that are time-tested or based on economic theory. Compare the SBC of the most reliable existing models and the best fitting model found from CDR.

Use the t-test to measure the statistical significance of each coefficient of the candidate model. All of the coefficients should have absolute t-values greater than 2.

If your model uses past behavior of the dependent variable as a leading indicator in the form of moving averages and autoregressions (ARMA terms), then you must run the Augmented Dickey-Fuller unit root test. This can provide protection against explosive gain from the moving average and autoregressive terms. (For details, refer to the macro %DFTEST.) Most likely, the best CDR models will employ both explanatory variables as well as ARMA terms. In this case, run the model without the ARMA terms and then run %DFTEST against the resulting residuals. If the %DFTEST indicates that the residuals are stationary, then the full model (which employs both the explanatory and ARMA terms) should be employed.

Collinearity poses computational problems often leading to spurious results. As an additional precaution, test for collinearity with the SINGULAR=1E-5 (and even 1E-4) and run candidate models with PROC MODEL to obtain collinearity diagnostics.

Many complex systems evolve over time. Older data may contain patterns inconsistent with more recent data. Economists often refer to such changes as “regime shifts” because of the radical changes in policy from one presidential or congressional regime to the next. While visualization techniques are more reliable, there are several worthwhile statistical tests. First, you can simply split the data into two consecutive halves and run PROC ARIMA on each of these data sets using the same variables and ARMA terms as found in the candidate model. Theoretically, the coefficients should be statistically equal. You can inspect the coefficients for yourself, or you can rely on formal significance tests such as the Chow test (which is an F-test) or the Watson-Davies.

Heteroscedacity occurs when the variance of the errors vary over time. Because regression assumes a constant variance, least squares estimates may have deceptively high significance tests. In PROC AUTOREG, specify the ARCHTEST option to detect this phenomena.

Further Detection of Possible Artifacts with Visualization

The mind’s eye can spot nuances and subtleties in the data that are often missed by statistical diagnostics because of the restrictive assumptions necessary to construct models and tests. Moreover, graphs appeal to the intuitive powers of the mind.

Outliers in the dependent and explanatory variables can bias all estimates and statistical tests. Because regression attempts to minimize the MSE, regression algorithms may inadvertently bias the fit to the outliers rather than to the underlying trend. Unfortunately, casual inspection of graphs of the raw data are not sufficiently intuitive for outlier detection.

I recommend filtering your data with the Median Smoothing technique. You then graph the filtered data as a curve with the raw data as points. Alternatively, you can subtract the raw data from the filtered data and graph the absolute values of their difference.

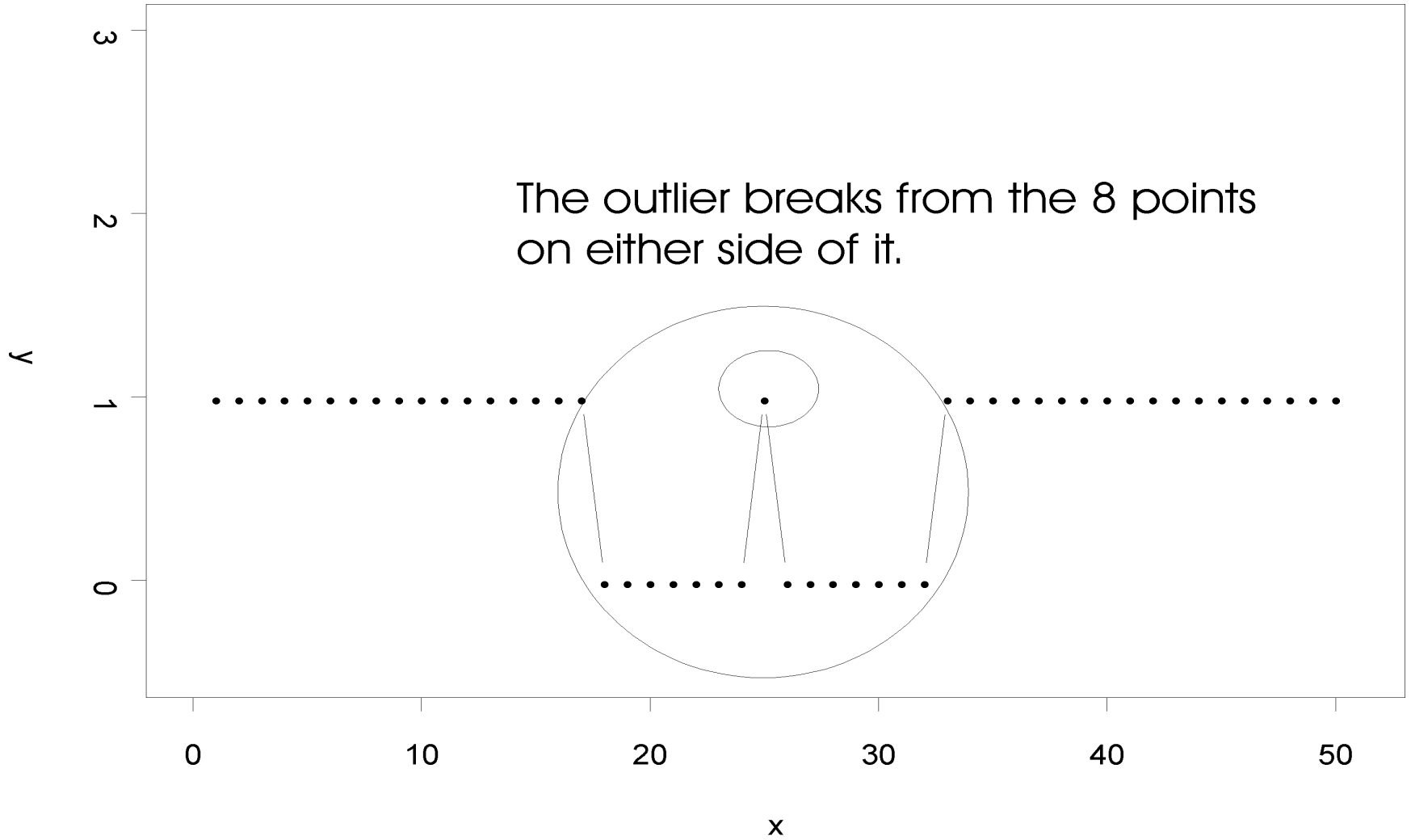
Median Smoothing detects a pattern for each neighborhood of points, pulling all of the deviant points in line with the local pattern. Most likely, median smoothing will eliminate severe noise while emphasizing the underlying pattern in the data. Unlike global representations of data, this method is not vulnerable to regime shifts, level shifts, or a series of outliers. Hence, these problems are more easily observed.

If severe outliers are detected, you should manually adjust outliers to more reasonable values and rerun the CDR on this “winsorized” data set. I do not recommend using impulse interventions because the SBC will penalize every outlier adjustment, possibly disqualifying a reasonable model for system behavior under normal conditions.

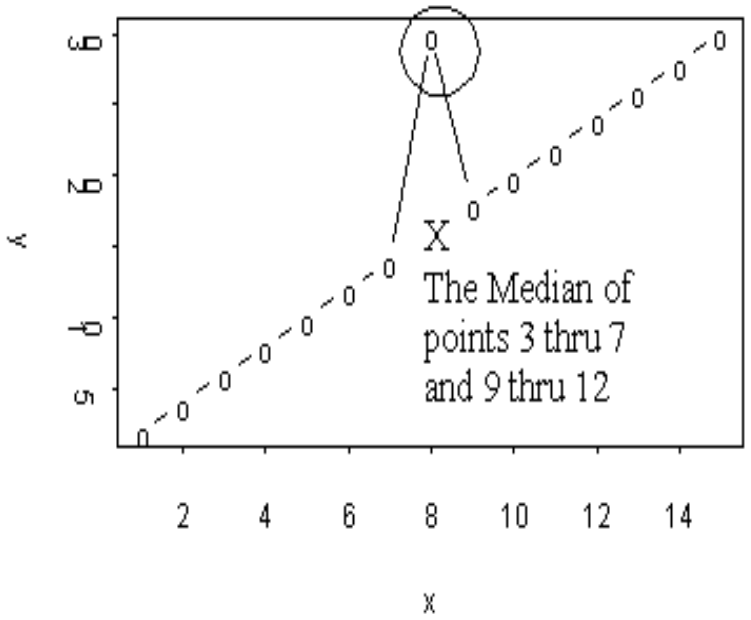
If regime shifts or level shifts are detected, you should employ intervention variables, as discussed in the section entitled “Intervention Models and Interrupted Time Series” in Chapter 3, "The ARIMA Procedure,” SAS/ETS User’s Guide, Version 6, Second Edition.

What follows is a visualization of the problem and my solution to the problem.

.



Outliers are an abrupt discontinuity when compared to the points around it. It is a local phenomona.



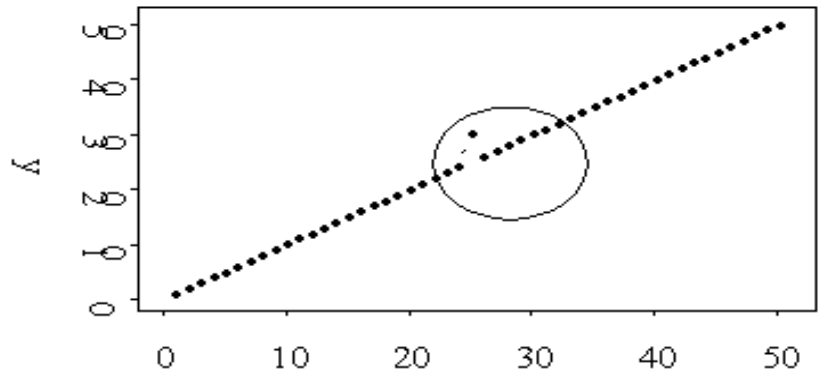
Because an outlier is an abrupt break with the pattern formed by its neighbors, we can often use the pattern formed by its neighbors to correct the outlier.

Fermat's local median bisection can correct outliers. For example, the outlier occurs at observation 8. We could take the 3 points preceding observation 8 and the 3 points after it and compute their median, which is 16, the correct number. The number of neighboring points we take from either side of the outlier is called its neighborhood and the number of points in this neighborhood (excluding the Outlier) divided by the number of points in the entire set is called the span. In this example, we are using a span of 6/15 or 40%. The smooth algorithm cross validates the median by comparing the local neighborhood median with different spans.

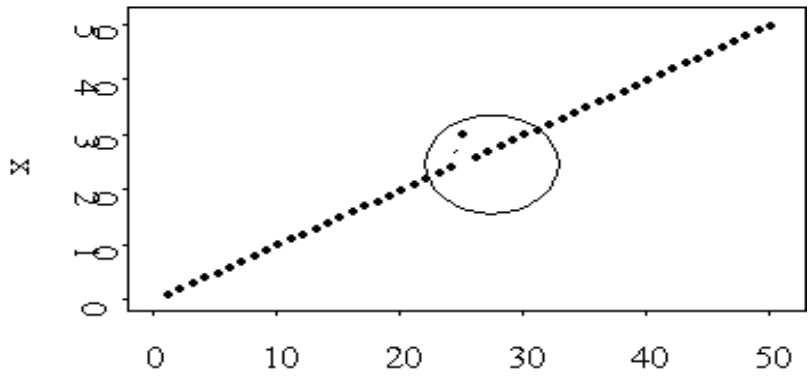
This simple method is very robust, that is, it can correct even large outliers automatically.

So, an outlier is a break in continuity with its neighbors and the smooth algorithm uses the pattern of its immediate neighbors to automatically correct this outlier.

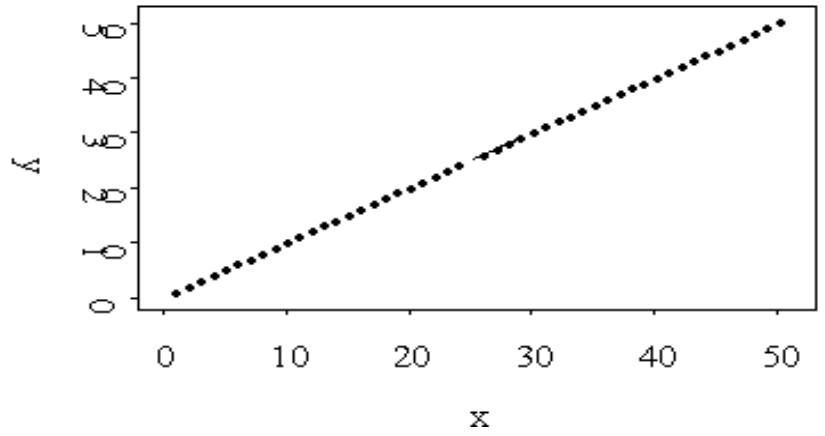
Handling Outliers: Here the Explanatory Variable, X, completely accounts for and explains the outlier at observation 25 in Y. That is, y as a function of x completely accounts for all of the points in Y including the outlier at observation 25.



Time series y has an outlier at 25



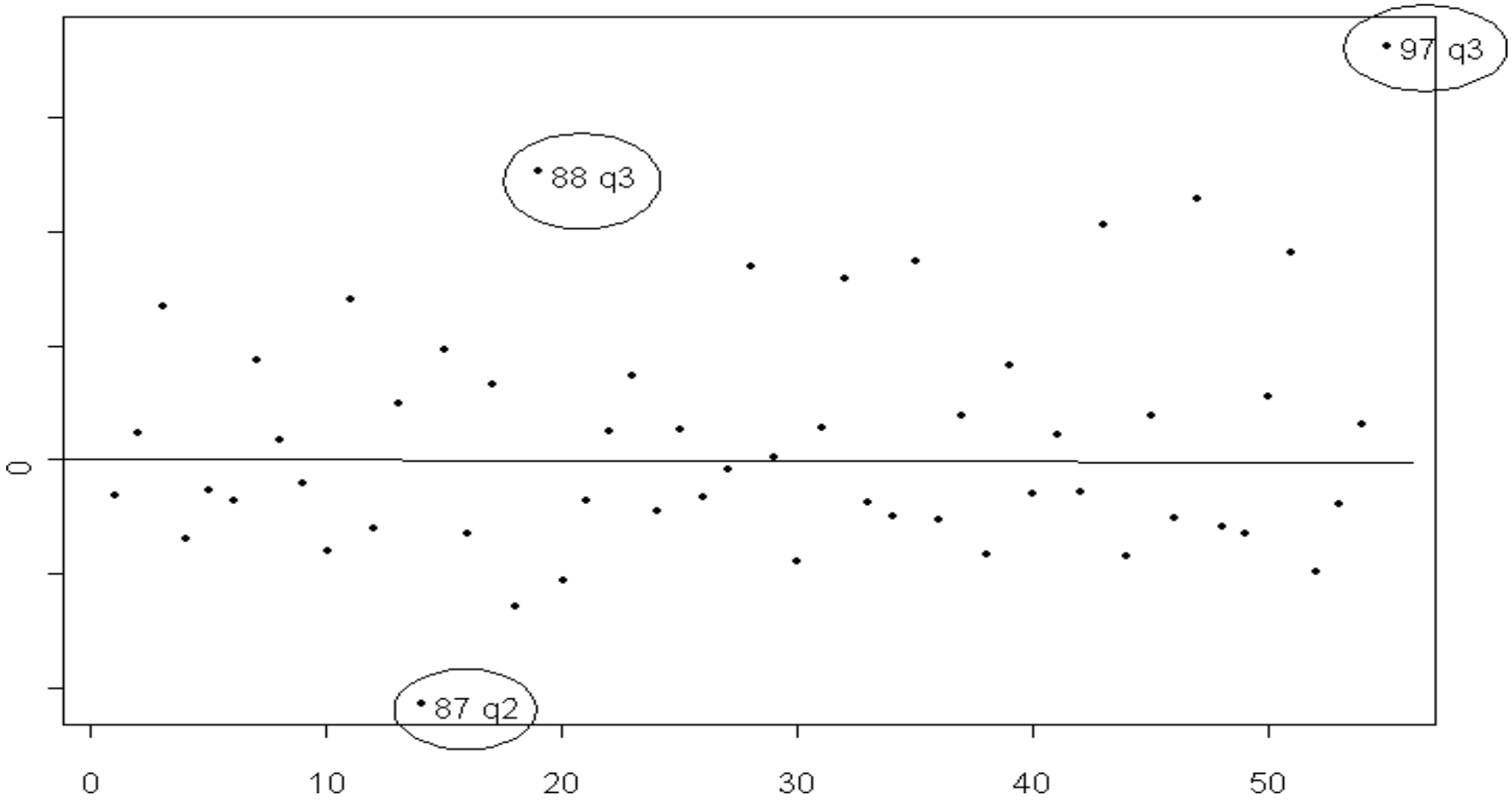
Time series x has an outlier at 25

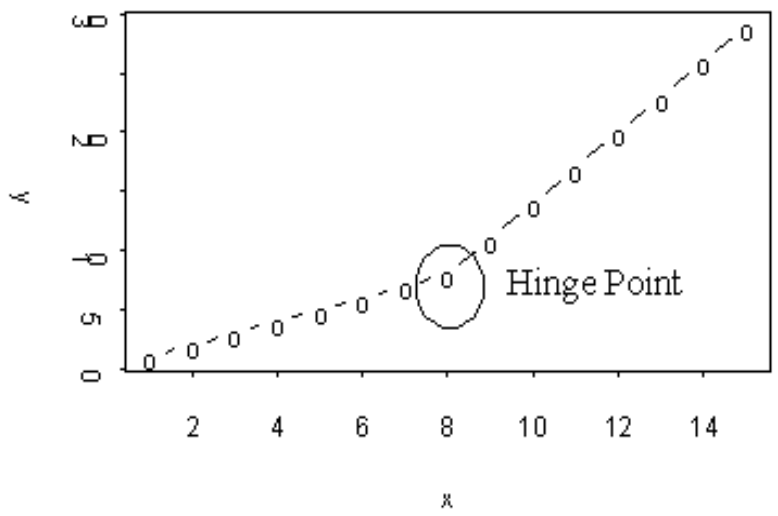


Y as a function of x,
in this case, y=x
the outlier at x completely explains the outlier of y and hence the model y= f(x) will not need any further adjustments.

Outliers Identified by Local Median Bisection Method

Since it is easier to spot outliers when the data is essentially horizontal, you can use the residuals of y - median smooth(y). The outliers are identified by date.





REGIME SHIFTS

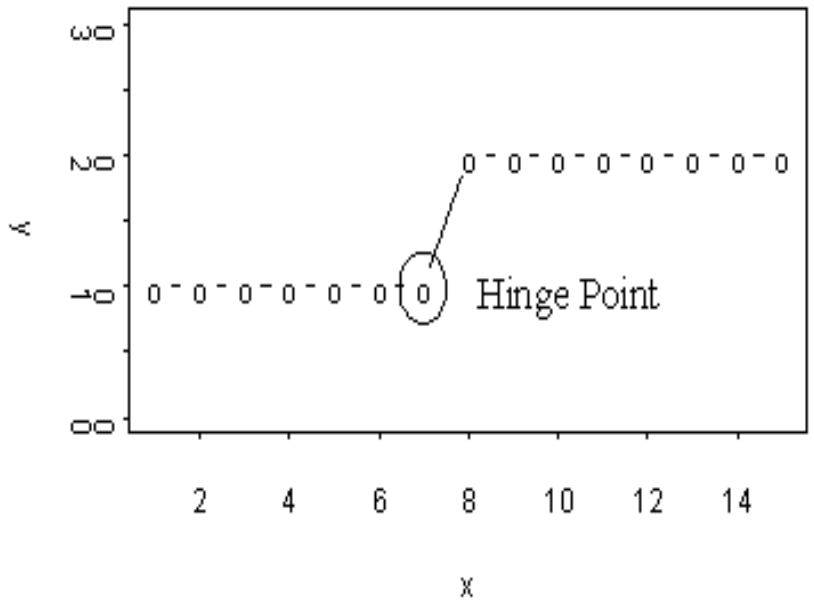
A Regime Shift represents a "permanant" and abrupt change in the pattern of points. An outlier is a temporary shift in the pattern of points formed by its neighbors. A Regime Shift is a departure from a previous pattern. The Pattern shifts at the hinge point. There is a slope of 1 before the hinge point at observation 8 and a slope of 2 after the hinge point.

While all 15 observations in the time series cannot be represented by a linear regression, the time series can be represented by 2 piece-wise linear regressions, one for each regime.

Note that each regime occurs over adjacent observations. In this example, Regime 1 occurs from observations 1 thru 7 and Regime 2 occurs from 8 thru 15. Each regime occurs over a sub-series, each regime occurs over contiguous subsets of x.

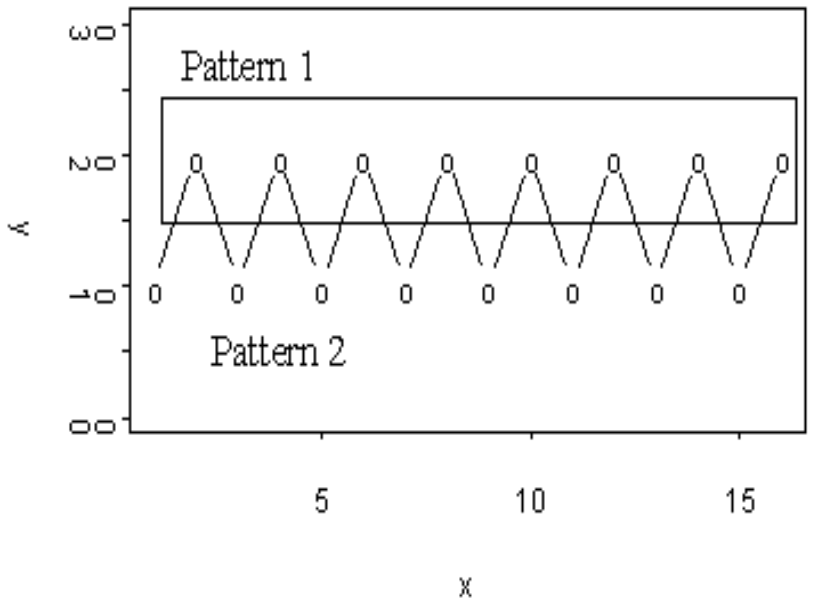
This appeals to our intuition because we assume that some underlying factors that influence y have changed radically. It could be a change in political regimes in Congress or a new President. Global warming may permanently effect the price of wheat or natural gas. An aging population has completely altered the investment strategies in the US financial markets. Thus, these changes occur over long periods of time and over contiguous sets of time.

Example 1



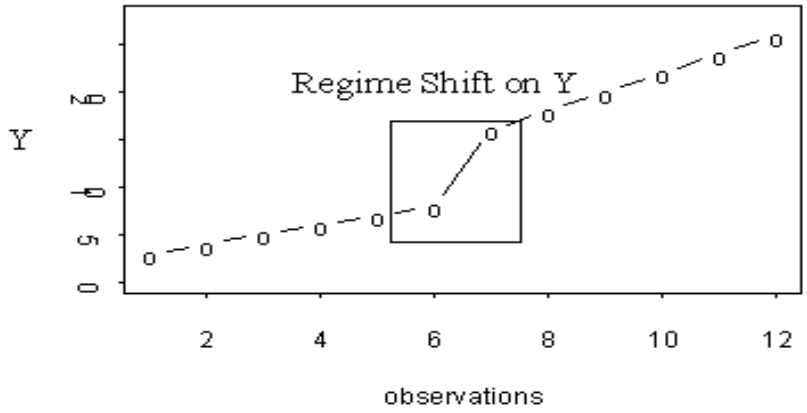
Example 1 represents a regime shift. In this case the intercept has changed from 1 to 2. Before the hinge point, all of the points are 1, that is, they form a consistent pattern over a contiguous subset of x, from 1 to 7. And after the hinge point, they form a different pattern, once again over a contiguous subset of x, from observations 8 to 16.

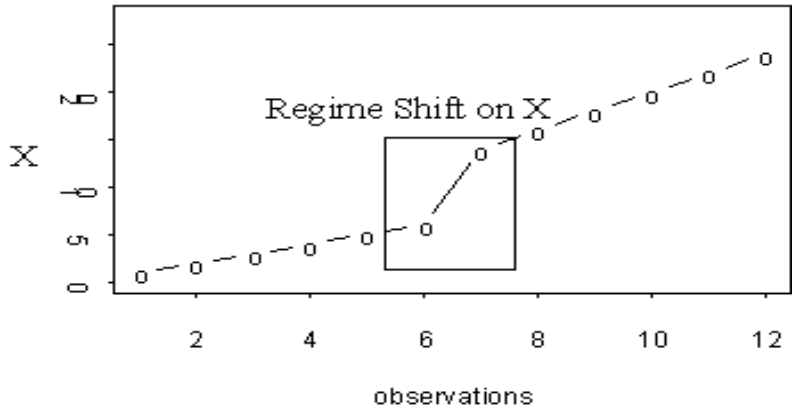
Example 2

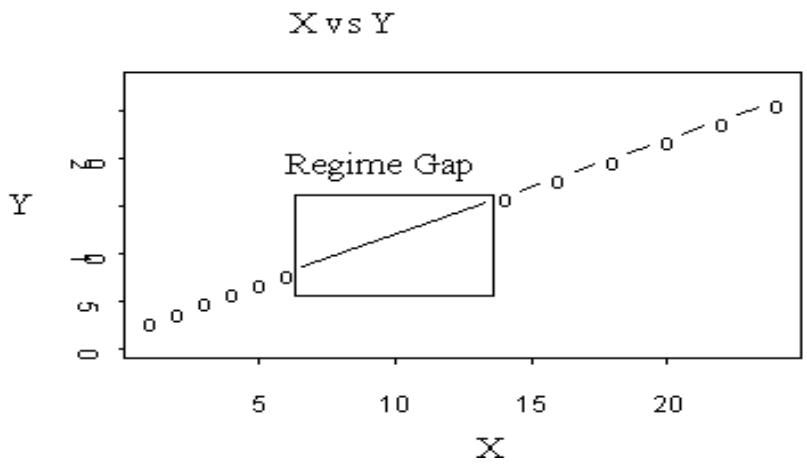


Example 2 is not a Regime Shift because its 2 patterns are interspersed over the observations. Pattern 1 occurs at all odd values and Pattern 2 occurs over all even values. But each pattern fails to occur over a contiguous subset of x.

In Dynamic Regression, we try to explain Y's behavior on the past behavior of Y and Explanatory Variables, often called "Leading Indicators". Since Dynamic Regression is a Multivariate Linear Model, Y is a Linear Function of X and variables created from Lags on Y. The model explains more of Y's behavior with a leading indicator rather than a dummy variable.





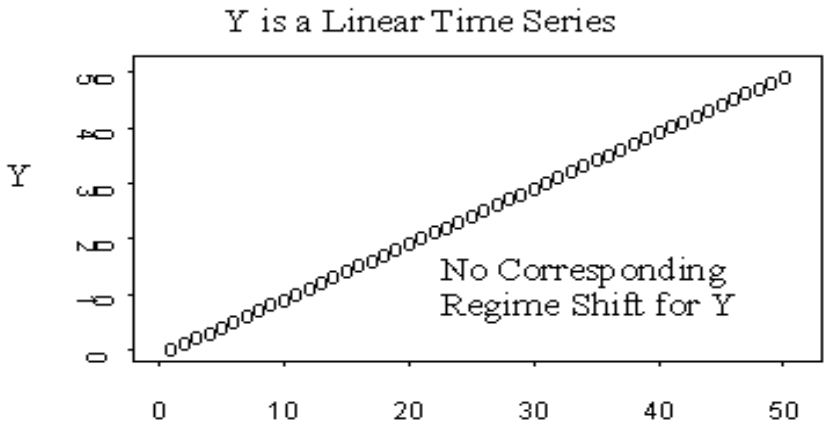


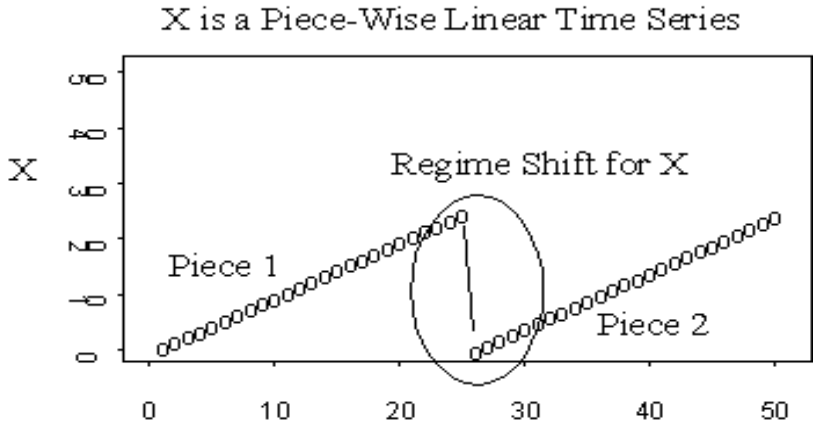
The Linear Regression of Y against X perfectly explains Y. Because the Regime Shifts in X and Y are correlated, X also explains the Regime Shift in Y. The Correlated Regime Shifts caused the Gap in the Graph of Y as a linear function of X but it also preserved linearity.

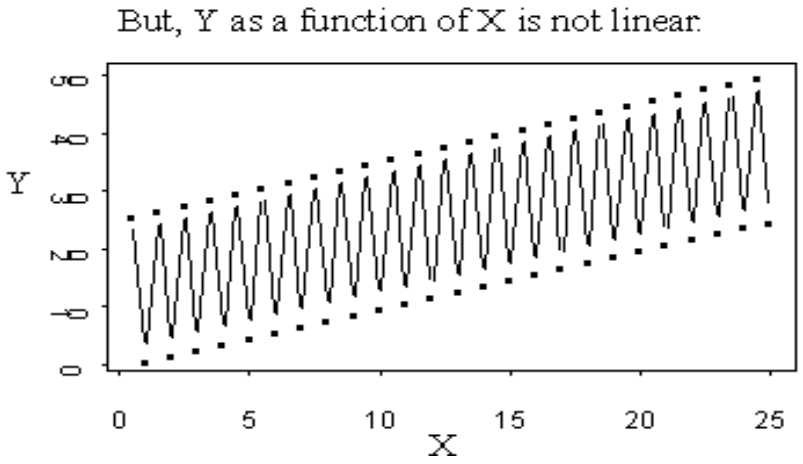
What would happen if the Regime Shifts do not match? Or, worse, if X had a Regime and Y did not?

In this paper, I hope to show how to handle this question.

This example shows the devastating power of distortion a regime shift in the explanatory variable X can have over y=f(x). A later section will show how to handle this nasty situation with differencing..







Y as a Time Series
with the column labeled T
representing time periods
1 thru 20

T	Y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	11
12	12
13	13
14	14
15	15
16	16
17	17
18	18
19	19
20	20

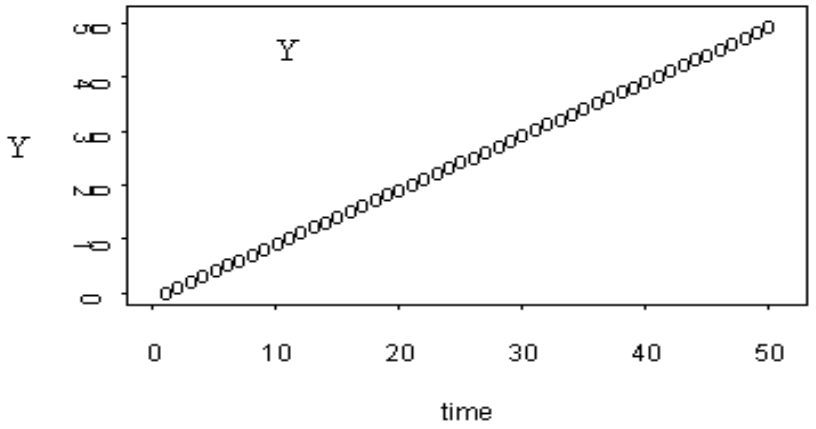
X as a Time Series
with the column labeled T
Note the break in the
pattern start with T=11

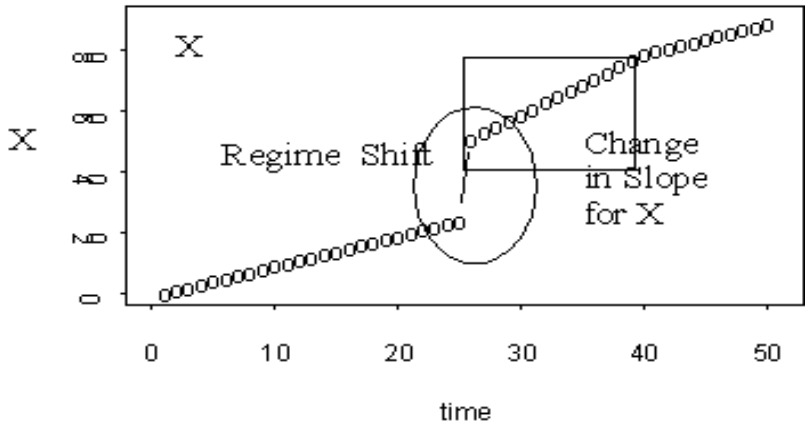
T	X
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10
11	0.5
12	1.5
13	2.5
14	3.5
15	4.5
16	5.5
17	6.5
18	7.5
19	8.5
20	9.5

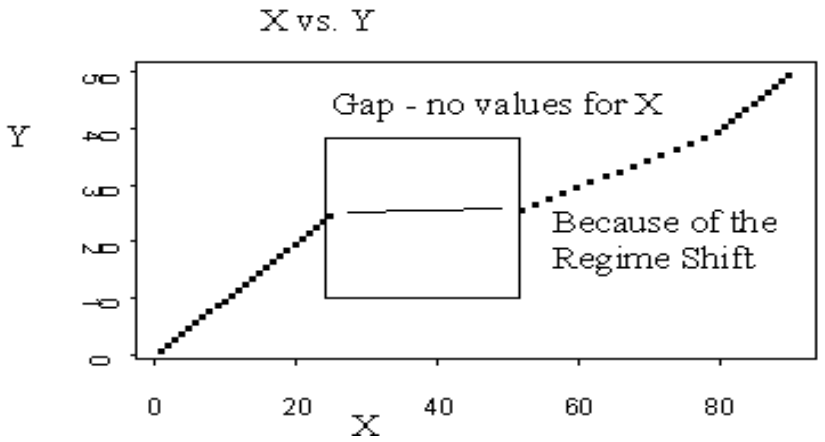
As a function of X,
both columns must be sorted
in ascending order on X

X	Y
0.5	11
1	1
1.5	12
2	2
2.5	13
3	3
3.5	14
4	4
4.5	15
5	5
5.5	16
6	6
6.5	17
7	7
7.5	18
8	8
8.5	19
9	9
9.5	20
10	10

A Regime Shift in X has a large discontinuity (a level shift) at time = 25. At that time, X jumps from 20 to 40. When Y is graphed against X, the pairs (X, Y) are sorted by the values of X. Thus, on the X vs Y graph, there is line connecting the points between X = 20 and X = 40. This gives the illusion that X vs Y is Non-Linear.



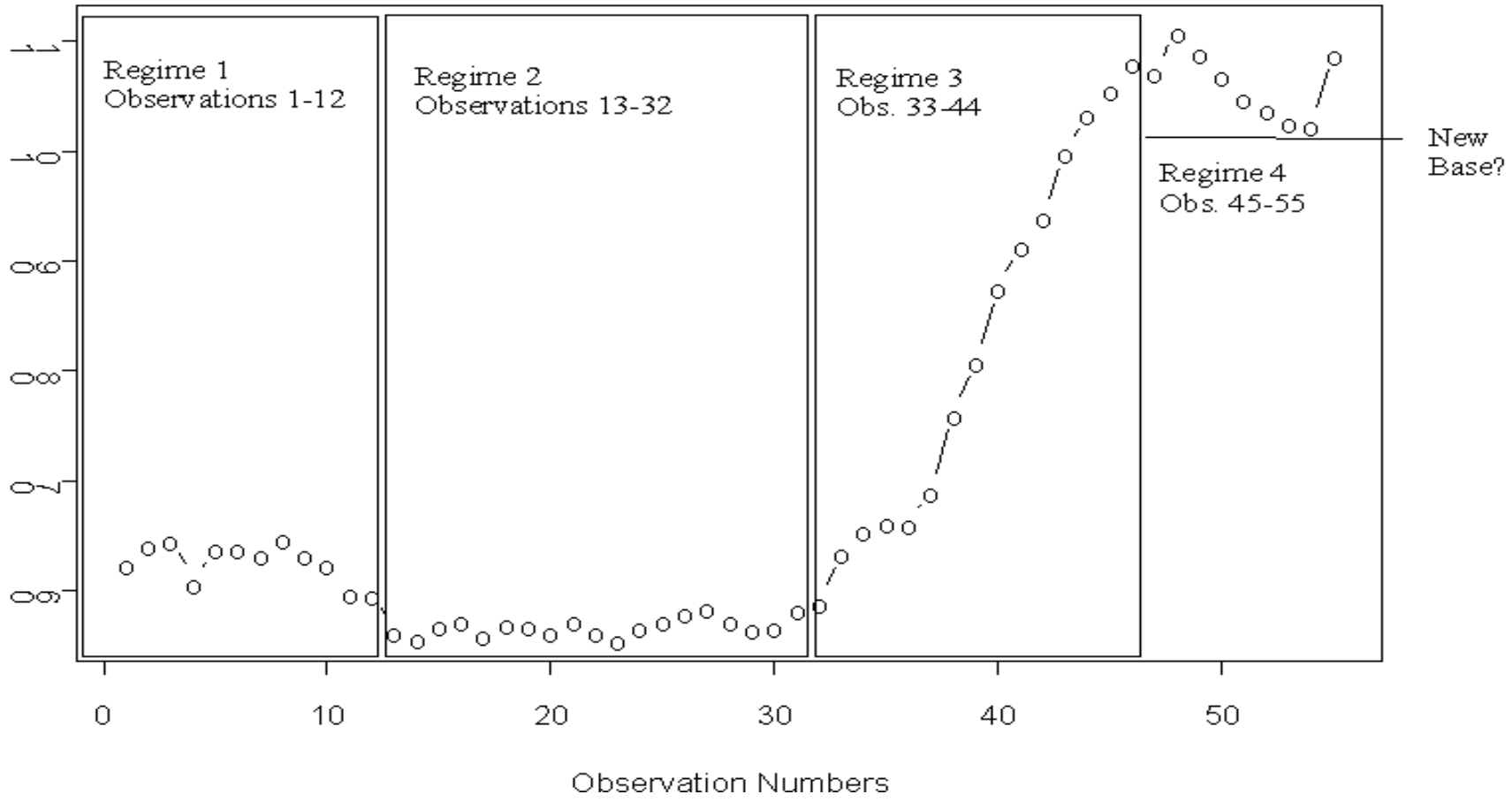




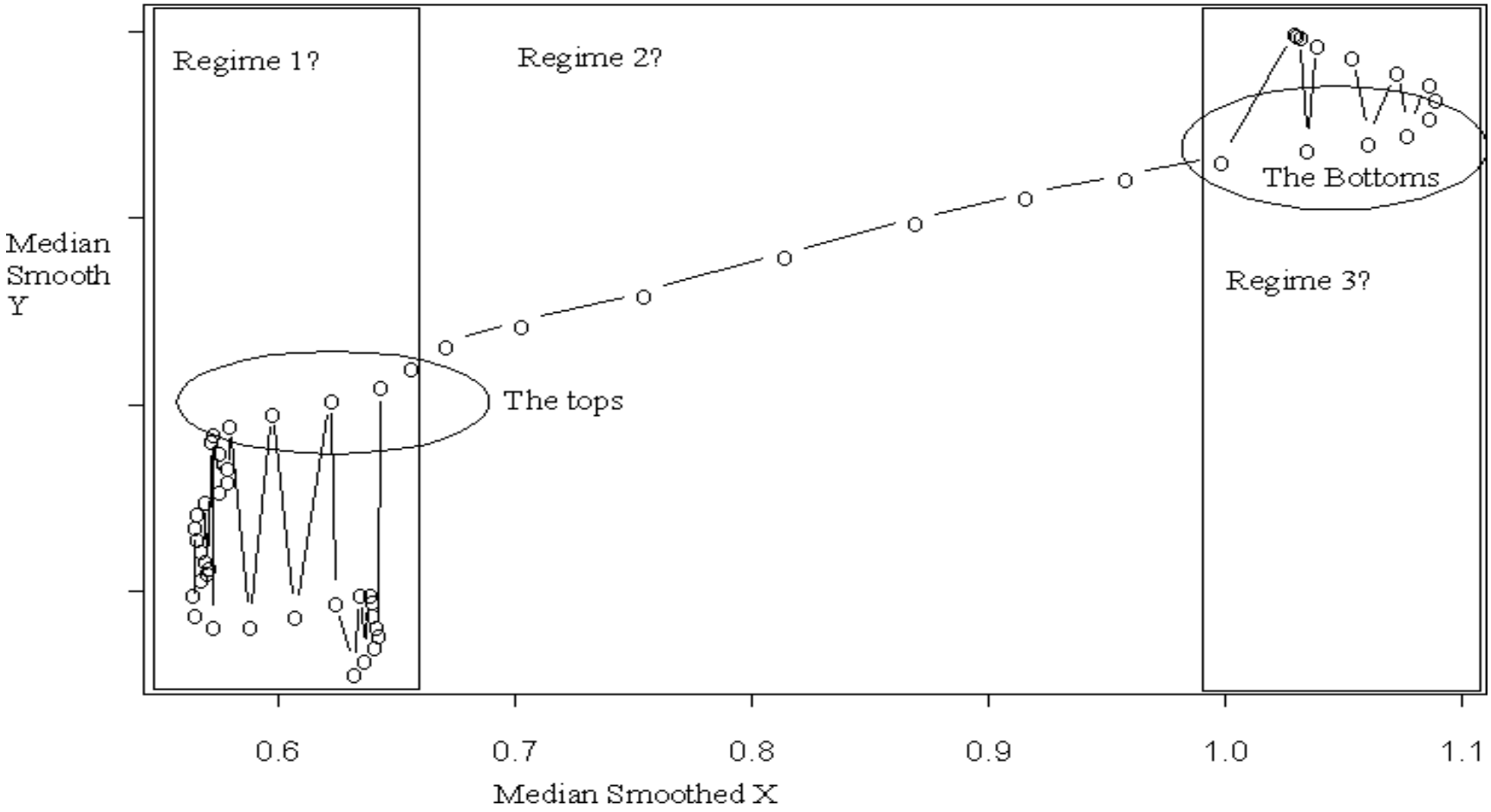
Y now appears non-linear because X controls the spacing between Y values.

In reality, X vs Y is piece-wise linear and we can still use dynamic piece-wise regession to construct a tight fitting model.

Here is the time series plot of median smoothed X. It has 4 Regimes with fuzzy boundaries.
We may want to combine Regimes 1 and 2.
In extrapolating beyond the historic data, the question becomes: where are we in regime 4.
From observations 1 thru 30 (60% of the time), X was relatively stable, forming almost a stationary wave.
However, from observation 32 to 44, X climbed steadily and sharply.



We cannot determine the Rregimes directly from the graph of X vs Y because X has many separate intervals of observations with nearly oscillating around the same level.
When we sort the pairs of (X,Y) in the historic data set by the values of X, we get the typical chaotic patterns. I attempted to arrive at a long term trend by selecting the tops of Regime 1 and the bottoms of Regime 3 without success. Neither the tops nor bottoms are from contiguous observations in the time-series of X.



Noise generated by outliers and level shifts frequently cause spurious models with deceptively lower MSE's than more representative models. That is, many artifacts model noise rather than the underlying pattern. Running models on median-smoothed, filtered data will eliminate many artifacts. Erroneous models that rely on exogenous changes for their strength will no longer have a unfair advantage over more reliable candidate models. In other words, we want to model the behavior of a system under normal or expected conditions rather than its response to occasional or random anomalies produced by influences outside of the system.

Often outliers and level shifts are not due to bad data. They often indicate the influence of variables not included in the systems analysis. Note the date of an outlier and try to determine what events may have triggered shocks to the system. Items in the news may alert you to abnormal behavior unpredictable by regression.

Also, PROC ARIMA describes the global or average responses of the system over a span of time within which data was collected. However, PROC ARIMA projects its forecasts based on the most recent data; that is, current data serves as a springboard for future estimates. Consequently, you should be particularly vigilant concerning recent abnormal behavior in the dependent and explanatory variables. Several outliers or anomalies in recent data may indicate a "turning point" where a system's dynamics shift radically, invalidating prior models based on older data generated by different systems dynamics.

Team Effort

Forecasting is an art best practiced as a team effort. If you know people who possess reliable knowledge acquired from years of study and experience on process behavior, then let them critique your new models. Dissent is healthy. It is the last line of defense against spurious models. CDR-based ARIMA forecasts are really a starting point for discussion, and the final forecast is arrived at by consensus. The human factor is critical.

Three Definitions Used in Combinatorics

Definition of a factor	A discrete, finite set of values describing the attribute of an entity or system. In demographics, the factor gender consists of the set {M, F}; . A factor in a CDR model is the possible lags on the dependent variable. In our example, this factor, the possible number of lags on the AR term, would consist of the set {0, 1, 2, 3, 4}; .
Definition of the cross product	An array formed from two or more factors. Each element of such an array is referred to as a "tuple". Each entry in a cross product is demarcated with " (" and ")". Each entry in a tuple is referred to as a component, and these entries are separated by "," with the first component corresponding to the first factor, the second component corresponding to the second factor, and so on.

Example of a cross product of two factors

In 1629, Fermat calculated the odds of "winning on field bets" at the game of dice by listing all possible combinations. Each die, d1 and d2, can assume values from 1 to 6. The cross product for a pair of dice contains 6x6=36 tuples.

Cross Product							Sum of the components within each tuple						
d1 \ d2	1	2	3	4	5	6							
1	(1, 1)	(1, 2)	*(1, 3)*	(1, 4)	(1, 5)	(1, 6);	(2)	(3)	***(4)***	(5)	(6)	(7)	
2	(2, 1)	*(2, 2)*	(2, 3)	(2, 4)	(2, 5)	(2, 6);	(3)	***(4)***	(5)	(6)	(7)	(8)	
3	*(3, 1)*	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6);	***(4)***	(5)	(6)	(7)	(8)	(9)	
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6);	(5)	(6)	(7)	(8)	(9)	(10)	
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6);	(6)	(7)	(8)	(9)	(10)	(11)	
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6);	(7)	(8)	(9)	(10)	(11)	(12)	

Fermat then evaluated the probability of a particular outcome by counting the number of occurrences of outcomes that sum to a particular number. For example, out of the 36 pairs, 3 pairs sum to a value of 4 (note the highlighted entries). Therefore, the odds of rolling a pair of dice that sum to 4 is 3/36 or 8.33%.

Outcome of d1+d2	Probability	Outcome of d1+d2	Probability
2	1/36	8	5/36
3	2/36	9	4/36
4	3/36	10	3/36
5	4/36	11	2/36
6	5/36	12	1/36
7	6/36		

The following DATA step will generate the cross product with two parameters: d1 is a factor = { 1 to 6} and is called the first component of the tuple. Likewise, d2 is a factor {1 to 6} and is called the second component of the tuple.

```
data dice(keep=d1 d2);
do d1=1 to 6;          /* iterate through all possible values of the first factor     */
do d2=1 to 6;          /* iterate through all possible values of the second factor */
output ;
end; end;
;
run;
```

Historic Footnote: In 1687, Jacob Bernoulli perfected Fermat=s original method. Wealthy Swiss and French gaming establishments paid for his research in order to change the rules of shooting dice to slightly favor the casino.

Definition of the combination	Unique groupings of the elements of a set; the number of elements allowed in each grouping is referred to as the Order of the Combination.

Example: Consider the set S = { a, b, c, d}. We will find all of set S′s combinations of order 2.

The SAS DATA step to generate the combinations from the set {a, b, c, d} would be as follows:

```
data c2(keep=combo);
array S[4] $1. ("a", "b", "c", "d");
do i =   1 to 4;
do j=    (i+1)  to 4;    /* j > i to prevent a duplicate selection made
previously by the loop on i */
do k=    (j+1)  to 4;    /* k > j to prevent a duplicate selection made
previously by the loop on j */
combo= "("  ||  compress(S[i])  ||  ", "  ||  compress(S[j])  ||   ", "  ||
 compress(S[k]) ||   ")";
output;
end; end; end;
;
run;
```

The results are
(a, b, c), (a, b, d), (a, c, d), (b, c, d)

The Building Blocks of Combinatorics

With the cross product and combination DATA steps (above) as building blocks, we will generate all of the ESTIMATE statements in PROC ARIMA. With simple variations of these two DATA steps, we will produce the list of all possible models.

Practical Examples: Doing a Forecast

Note: We have taken logs of all variables in the data set lyxx.

QTR	ATGR	WSAG	WSCONS	WSFIRE	WSGOV	WSMANU	WSMIN	WSTCU	WSSERV	WSRETL	WSWHL	DEFL
84Q1	21.993	-2.659	-0.476	-0.942	0.990	-0.448	-0.569	-0.327	0.541	-0.066	-0.889	4.280
84Q2	22.027	-2.645	-0.448	-0.919	0.997	-0.416	-0.557	-0.296	0.582	-0.027	-0.851	4.289
84Q3	22.075	-2.631	-0.440	-0.879	1.019	-0.387	-0.534	-0.285	0.619	-0.013	-0.828	4.296
84Q4	22.038	-2.631	-0.499	-0.860	1.028	-0.380	-0.528	-0.284	0.636	0.005	-0.810	4.304
85Q1	22.062	-2.604	-0.445	-0.816	1.061	-0.340	-0.526	-0.276	0.651	0.018	-0.792	4.314
85Q2	22.077	-2.590	-0.440	-0.792	1.063	-0.326	-0.516	-0.270	0.682	0.027	-0.770	4.324
85Q3	22.126	-2.590	-0.448	-0.774	1.085	-0.311	-0.541	-0.260	0.691	0.058	-0.774	4.333
85Q4	22.119	-2.604	-0.425	-0.757	1.090	-0.293	-0.562	-0.255	0.719	0.075	-0.761	4.343
86Q1	22.108	-2.617	-0.448	-0.726	1.111	-0.281	-0.594	-0.259	0.747	0.085	-0.774	4.350
86Q2	22.081	-2.645	-0.459	-0.724	1.103	-0.281	-0.742	-0.259	0.756	0.089	-0.790	4.351
86Q3	22.119	-2.645	-0.503	-0.701	1.122	-0.265	-0.849	-0.274	0.777	0.087	-0.790	4.359
86Q4	22.055	-2.645	-0.506	-0.673	1.128	-0.263	-0.870	-0.264	0.801	0.102	-0.787	4.367
87Q1	22.081	-2.645	-0.564	-0.681	1.176	-0.278	-0.906	-0.254	0.808	0.110	-0.761	4.379
87Q2	22.030	-2.617	-0.576	-0.660	1.166	-0.250	-0.842	-0.246	0.839	0.123	-0.749	4.388
87Q3	22.138	-2.577	-0.555	-0.658	1.171	-0.227	-0.856	-0.242	0.857	0.151	-0.726	4.399
87Q4	22.118	-2.551	-0.546	-0.662	1.186	-0.188	-0.805	-0.223	0.909	0.148	-0.705	4.410
88Q1	22.158	-2.513	-0.587	-0.713	1.202	-0.174	-0.810	-0.248	0.890	0.162	-0.717	4.416
88Q2	22.117	-2.501	-0.560	-0.650	1.209	-0.150	-0.787	-0.214	0.948	0.168	-0.675	4.428
88Q3	22.216	-2.489	-0.553	-0.625	1.177	-0.135	-0.823	-0.208	0.976	0.187	-0.658	4.441
88Q4	22.146	-2.465	-0.553	-0.635	1.216	-0.092	-0.832	-0.203	0.981	0.196	-0.641	4.453

Note: Where QTR means the fiscal quarter, the suffix WS stands for Wages and Salaries and ATGR (the response variable) means Adjusted Taxable Gross Receipts.

All of the data used in these examples is derived from the last 55 quarters of the New Mexico economy. Because much of the state=s economy (as measured in dollars) grew exponentially from inflation and population growth, the log transformation is appropriate. Also, in ten years of testing, the log transform has performed well.

Establishing a Benchmark for Model Accuracy: The Univariate BJ ARIMA Model

Because of its simplicity and elegance, every CDR candidate model must exceed the accuracy of the univariate BJ ARIMA. The BJ ARIMA model relies entirely on past behavior to forecast future behavior of the dependent variable. After 20 years, it is still accepted as the benchmark modeling paradigm.

The univariate ARIMA will help constrain the number of possible AR and MA values in the dynamic regressions. Better constraints will reduce the number of combinations that PROC ARIMA must evaluate.

If the dynamic regression uses an ARIMA component, particularly an MA component, then the dependent variable must be cointegrated with the explanatory variables (and their lags).

Example of a univariate BJ ARIMA
Adjusted Taxable Gross Receipts resembles a sales tax with an extremely broad base. It applies to almost all final sales at the end of the chain of commerce (that is, it is not a pyramid tax). We want to model the ATGR based solely on its past behavior.

First, we must apply transforms that make ATGR stationary. In particular, the mean and all statistics based on the second moment must be invariant over time. We applied a log to ATGR .

```
proc arima data=lyxx;
i var=atgr(1) stationarity=(adf=(4));
run;
```

ARIMA Procedure

Name of variable = ATGR.
Period(s) of Differencing = 1.

Autocorrelations

```
       Lag Covariance Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
Std
         0  0.0013486   1.00000     |                  |********************|
 0
         1 -0.0007685   -0.56981    |        **********|   .                |
 0.134840
         2 0.00042225    0.31310    |             .    |******.            |
 0.173171
         3 -0.0005321   -0.39454    |          ********|      .             |
 0.183175
         4 0.00077771    0.57667  |            .       |************        |
  0.198024
         5 -0.0006342   -0.47025  |          *********|        .           |
  0.226509
         6 0.00036799    0.27286  |          .         |*****    .          |
  0.243613
         7 -0.0004354   -0.32285  |          .   ******|         .          |
  0.249108
         8 0.00047701    0.35370  |          .         |*******  .          |
  0.256603
         9 -0.0003097   -0.22964  |         .     *****|          .         |
  0.265319
        10 0.00015593    0.11562  |         .          |**        .         |
  0.268909
        11 -0.0003072   -0.22778  |         .     *****|          .         |
  0.269811
        12 0.00031396    0.23280  |         .          |*****     .         |
  0.273285
        13 -0.0000974   -0.07221  |         .         *|          .         |
  0.276867
                                              "." marks two standard errors
```

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.32703	|. |******* |
2	0.03109	|. |* . |
3	-0.08498	|. ** | . |
4	-0.16576	|. *** | . |
5	0.19132	|. |****. |
6	0.08843	|. |** . |
7	0.03745	|. |* . |
8	-0.02444	|. | . |
9	0.01552	|. | . |
10	0.13215	|. |*** . |
11	0.11115	|. |** . |
12	-0.00133	|. | . |
13	-0.05662	|. * | . |

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	-0.56981	| ********** | . |
2	-0.01715	|. | . |
3	-0.33008	| ******* | . |
4	0.37063	|. |******* |
5	-0.01629	|. | . |
6	-0.06499	|. * | . |
7	-0.11873	|. ** | . |
8	-0.10362	|. ** | . |
9	0.15615	|. |*** . |
10	-0.06627	|. * | . |
11	-0.12439	|. ** | . |
12	-0.07421	|. * | . |
13	0.08171	|. |** . |

* The White Noise is too high without accounting for seasonality.

Autocorrelation Check for White Noise

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	73.10	6	0.000	-0.570	0.313	-0.395	0.577	-0.470	0.273
12	100.43	12	0.000	-0.323	0.354	-0.230	0.116	-0.228	0.233

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	RHO	Prob<RHO	T	Prob<T	F	Prob<F
Zero Mean	4	-5.3888	0.1050	-1.4060	0.1466	--	--
**Single Mean	4	-36.5771	0.0004	-2.7656	0.0704	3.8283	0.1272
Trend	4	-48.8250	0.0001	-2.9338	0.1610	4.3079	0.3457

.

** Without the dummy variable to account for seasonality, this ADF result is close enough.

The ADF results suggest var = a tgr(1). The acf chart decays exponentially to insignificance at about the fifth lag. And the PACF cuts off after two lags. It looks like an AR(4) process. Seasonality is handled with a dummy variable rather than a deterministic dif = (1,4) or dif = (1)(4).

A standard model with a single difference, an intervention dummy variable DUMQ3 (which has a 1 at the third quarter of every year and 0 otherwise) produces good working results.

Estimate p = (1,4) input = (dumq3);

Parameter	Estimate	Approx. Std Error	T Ratio	Lag	Variable	Shift
MU	0.0028497	0.0038104	0.75	0	ATGR	0
AR1,1	-0.46379	0.11831	-3.92	1	ATGR	0
AR1,2	0.25118	0.11833	2.12	4	ATGR	0
NUM1	0.04014	0.01031	3.89	0	DUMQ3	0

Strong T-Tests for the AR terms and the dummy seasonal variable

Constant Estimate = 0.0034556

Variance Estimate = 0.00061439
Std Error Estimate = 0.02478688
AIC = -246.78816*
* Benchmark SBC = -238.75883* *The BJ ARIMA benchmark SBC*
Number of Residuals= 55
* Does not include log determinant.

Correlations of the Estimates

Variable	Parameter	ATGR MU	ATGR AR1,1	ATGR AR1,2	DUMQ3 NUM1
ATGR	MU	1.000	0.044	-0.013	-0.697
ATGR	AR1,1	0.044	1.000	0.231	-0.070
ATGR	AR1,2	-0.013	0.231	1.000	0.023
DUMQ3	NUM1	-0.697	-0.070	0.023	1.000

These are excellent cross correlations between the parameters.

Autocorrelation Check of Residuals

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	2.80	4	0.591	0.029	0.003	0.044	-0.025	-0.199	0.042

This model=s residuals are close to white noise. This is an acceptable benchmark BJ ARIMA model.

Running the Combinatoric Dynamic Regression

We want to model adjusted gross receipts taxes (ATGR) as a function of only two Sectors. In this model, we will rely upon only the major revenue-producing sectors in this state's economy:

List of W&S (Wages and Salaries) Sectoral Explanatory Variables:

wscons	W&S Construction	wsfire	W&S Finance, Insurance, Real Estate
wsmanu	W&S Manufacturing	wsmin	W&S Mining
wsretal	W&S Retail Sales	wsserv	W&S Service
wstcu	W&S Transportation, Communications and Utilities		

with two factors appearing in all models (hence their inclusion will not increase the number of combinations):

defl	GDP Deflator
dumq3	a seasonal intervention for the strongest, the third quarter of every year.

Note: W&S means Sectoral Wages and Salaries earned in New Mexico.

We want a pair of lags, one for each explanatory variable, that produces the "best" fit as measured by the SBC.

Format of the Model: ESTIMATE input = (lag1$ X1 lag2$ X2 DEFL DUMQ3);

Constraints:

2 W&S sectoral explanatory variables (denoted by X1 and X2) must be used in each model.
X1 and X2 can have a single lag varying from 0 to 5 quarters.
Each lagged cross correlation should exceed 0.65.

Number of Models in the List of Combinations:

The number of combinations of 2 distinct sectors drawn from a set of 7 sectors is	21.
The size of the Cross Product of the 2 lags is 6 x 6	36.
Total Number (1) x (2)	756.

Generate the List of Combinations of Models (The MOD File)

```
data mod(keep=est);
array vv[9]   $8.
(Awscons","wsfire","wsmanu","wsmin","wsretl","wsserv","wstcu","defl","dumq3");
array sym[6] $5. (A @, "1", "2", "3", A4", A5");
retain modelno 0;
/* Combination Phase -  select a pair of W&S explanatory variables      */
do v1=1 to 7; /* select the first wage and salary explanatory variable  */
do v2=(v1+1) to 7;  /* select the second unique wage and salary explanatory */
                    /* variable                                          */

/* Construct a Cross Product for a pair of lags                         */
do s1=1 to 6;     /* select a lag for the first explanatory variable    */
do s2=1 to 6;     /* select a lag for the second explanatory variable   */
modelno=modelno+1;  /* assign a Model Identification Number             */
est=compress(@mod@ || modelno)||@:  A||"estimate  input=("
|| compress(sym[s1])  ||  @  A  ||  compress(vv[v1])||" " || compress(sym[s2])
||  @  A  ||  compress(vv[v2]) ||  "  defl dumq3);";
output ;
end; end; end; end;
;
run;
```

```
Here is a partial list of combinations (out=mod):

   mod1:    estimate  input=(    wscons     wsfire    defl dumq3);
   mod2:    estimate  input=(    wscons 1   wsfire    defl dumq3);
   mod3:    estimate  input=(    wscons 2   wsfire    defl dumq3);
   mod4:    estimate  input=(    wscons 3   wsfire    defl dumq3);
   mod5:    estimate  input=(    wscons 4   wsfire    defl dumq3);
   mod6:    estimate  input=(    wscons 5   wsfire    defl dumq3);
   mod7:    estimate  input=(1   wscons     wsfire    defl dumq3);
   mod8:    estimate  input=(1   wscons 1   wsfire    defl dumq3);
   mod9:    estimate  input=(1   wscons 2   wsfire    defl dumq3);
   mod10:   estimate  input=(1   wscons 3   wsfire    defl dumq3);
   mod11:   estimate  input=(1   wscons 4   wsfire    defl dumq3);
   mod12:   estimate  input=(1   wscons 5   wsfire    defl dumq3);
     ooo              ooo            ooo
   mod250:  estimate  input=(5   wsfire 3    wsmanu    defl dumq3);
   mod251:  estimate  input=(5   wsfire 4    wsmanu    defl dumq3);
   mod252:  estimate  input=(5   wsfire 5    wsmanu    defl dumq3);
   mod253:  estimate  input=(    wsfire      wsmin     defl dumq3);
   mod254:  estimate  input=(    wsfire 1    wsmin     defl dumq3);
   mod255:  estimate  input=(    wsfire 2    wsmin     defl dumq3);
   mod256:  estimate  input=(    wsfire 3    wsmin     defl dumq3);
   mod257:  estimate  input=(    wsfire 4    wsmin     defl dumq3);
     ooo             ooo             ooo
   mod746:  estimate  input=(4   wsserv 1    wstcu     defl dumq3);
   mod747:  estimate  input=(4   wsserv 2    wstcu     defl dumq3);
   mod748:  estimate  input=(4   wsserv 3    wstcu     defl dumq3);
   mod749:  estimate  input=(4   wsserv 4    wstcu     defl dumq3);
   mod750:  estimate  input=(4   wsserv 5    wstcu     defl dumq3);
   mod751:  estimate  input=(5   wsserv      wstcu     defl dumq3);
   mod752:  estimate  input=(5   wsserv 1    wstcu     defl dumq3);
   mod753:  estimate  input=(5   wsserv 2    wstcu     defl dumq3);
   mod754:  estimate  input=(5   wsserv 3    wstcu     defl dumq3);
   mod755:  estimate  input=(5   wsserv 4    wstcu     defl dumq3);
   mod756:  estimate  input=(5   wsserv 5    wstcu     defl dumq3);
```

Save SAS/ETS Statements

Next, export the SAS data set MOD to a text file.
Open this file in a word processor and copy the MOD text onto the clipboard.

Use the SAS/ETS program below to evaluate all of the models (ESTIMATE statements):

```
filename newout "d:\1a\fermat.txt";
proc printto print=newout new;
proc arima data=lyxx;
identify var=atgr crosscorr=(WSCONS WSFIRE WSMANU WSMIN WSRETL WSSERV WSTCU
  DEFL DUMQ3) noprint;

/* next, paste the clipboard text from the word processor here     */
;
run;
proc printto;
run;
```

```
/* extract the model number and resulting SBC from PROC ARIMA output     */
data xx(keep=modelno SBC);
retain modelno 0;
infile "d:\1a\fermat.txt";
input zz $97.;
ii=index(zz,"SBC");
if (ii NE 0) then do;
pp=index(zz,"=")+1;
SBC=input(substr(zz,pp),11.);
modelno=modelno+1;
output;
end;
;
proc sort data=xx out=rank;
by SBC;
;
proc print data=rank;
run;
```

Here are the top 10 models ranked by SBC as a measure of goodness of fit:

```
MODELNO     SBC   Corresponding Estimate Statements

 80   -269.442    estimate  input=(1      wscons  1  wsmin     defl dumq3);
 79   -266.886    estimate  input=(1      wscons     wsmin     defl dumq3);
 73   -266.462    estimate  input=(       wscons     wsmin     defl dumq3);
 74   -265.581    estimate  input=(       wscons   1 wsmin     defl dumq3);
182   -263.801    estimate  input=(       wscons   1 wstcu     defl dumq3);
188   -262.899    estimate  input=(1      wscons   1 wstcu     defl dumq3);
181   -262.427    estimate  input=(       wscons     wstcu     defl dumq3);
 37   -262.055    estimate  input=(       wscons     wsmanu    defl dumq3);
109   -260.924    estimate  input=(       wscons     wsretl    defl dumq3);
  1   -260.116    estimate  input=(       wscons     wsfire    defl dumq3);
```

Find the ESTIMATE statement corresponding to a model number by looking it up in the MOD.TXT file.

Find the ARIMA listing corresponding to a model number by looking it up in the FERMAT.TXT file.

Validate Candidate Model - ID 80

Parameter	Estimate	Std Error	**T Ratio**	Lag	Variable	Shift
MU	17.38765	0.12331	**141.01**	0	ATGR	0
NUM1	0.26704	0.01864	**14.33**	0	WSCONS	1
NUM2	0.10670	0.02799	**3.81**	0	WSMIN	1
NUM3	1.12523	0.02680	**41.98**	0	DEFL	0
NUM4	0.02969	0.00566	**5.24**	0	DUMQ3	0

We note that the T-ratios are highly significant at an alpha < .01!

The correlations between the coefficients are low.

Correlations of the Estimates

Variable	ATGR	WSCONS	WSMIN	DEFL	DUMQ3
ATGR	1.000	0.631	-0.075	-0.992	-0.032
WSCONS	0.631	**1.000**	**-0.637**	**-0.683**	**0.007**
WSMIN	-0.075	**-0.637**	**1.000**	**0.192**	**-0.046**
DEFL	-0.992	**-0.683**	**0.192**	**1.000**	**0.014**
DUMQ3	-0.032	**0.007**	**-0.046**	**0.014**	**1.000**

The Ljung-Box Q Test on the residuals reveals that this model sufficiently accounts for most of the variation in ATGR. However, lags 1, 3, and particularly 5 are weak. This model could use some fine tuning.

Autocorrelation Check of Residuals

To Lag	Chi Square	DF	Prob	Autocorrelations					
6	10.26	6	0.114	0.010	0.190	-0.023	0.137	-0.304	-0.132
12	19.40	12	0.079	-0.190	-0.173	-0.075	-0.185	-0.068	-0.152

The ACF check of the residuals indicate that there may be some information to be accounted for by a more complex model. Fine-tuning is beyond the scope of this article.

We begin checking for model adequacy.

```
proc arima data=lyxx;
identify var=atgr crosscorr=(wscons wsmin dumq3 defl) noprint;
est input=(1 wscons 1 wsmin dumq3 defl) plot;
run;
```

There is a problem caused by an unspecified moving average on the error term. The inverse autocorrelations are all positive and do not decay quickly enough.

Inverse Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.20386	|. |****. |
2	0.13956	|. |*** . |
3	0.12091	|. |** . |
4	0.06506	|. |* . |
5	0.40142	|. |******** |
6	0.26042	|. |***** |
7	0.22259	|. |****. |
8	0.11367	|. |** . |
9	0.05503	|. |* . |
10	0.13732	|. |*** . |
11	0.14161	|. |*** . |
12	0.15825	|. |*** . |
13	0.01642	|. | . |
14	0.09984	|. |** . |

And worse, the PACF does not cut off or decay exponentially. Furthermore, it is negative most of the time. The PACF demonstrates some inadequacy in this model. Adding an MA term to the model might produce improvement; however, such fine-tuning is beyond the scope of this paper.

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.00975	|. | . |
2	0.18993	|. |****. |
3	-0.02696	|. *| . |
4	0.10582	|. |** . |
5	-0.31337	|****** | . |
6	-0.18254	|.**** | . |
7	-0.09286	|. ** | . |
8	-0.16375	|. *** | . |
9	0.05742	|. |* . |
10	-0.22473	|.**** | . |
11	-0.15668	|. *** | . |
12	-0.23036	|***** | . |
13	0.00183	|. | . |
14	-0.14016	|. *** | . |

Comparison of CDR Explanatory Model and the BJ ARIMA

The measures of goodness of fit look promising, producing roughly a 13% improvement for the SBC and a 25% improvement for the Std Error Estimate over the univariate ARIMA.

	CDR	BJ ARIMA	
Variance Estimate	= 0.00033352	0.00061	
Std Error Estimate	= 0.01826247	0.02478	CDR shows a 25% improvement!
AIC	= -279.47	-246.7	
SBC	= -269.44	-238.7	CDR shows a 13% improvement!

with Number of Residuals = 55

BACK-CAST for Univariate BJ ARIMA

Obs	Forecast	Std Error	Lower 95%	Upper 95%	Actual	Residual
53	22.6693	0.0248	22.6207	22.7179	22.6633	-0.0060
54	22.6703	0.0281	22.6152	22.7255	22.6749	0.0045
55	22.7106	0.0337	22.6445	22.7767	22.7215	0.0109
56	22.7064	0.0374	22.6331	22.7797	22.7112	0.0048

Back-Casts for CDR Model 80: The Std Error shows a 34% improvement over the BJ ARIMA!

Obs	Forecast	Std Error	Lower 95%	Upper 95%	Actual	Residual
53	22.6562	0.0183	22.6204	22.6920	22.6633	0.0071
54	22.6581	0.0183	22.6223	22.6939	22.6749	0.0168
55	22.7205	0.0183	22.6847	22.7563	22.7215	0.0010
56	22.6992	0.0183	22.6634	22.7350	22.7112	0.0120

In general, the CDR model back fits the last four quarters as well as the BJ ARIMA. The CDR candidate shows tremendous improvement in the confidence interval. However, the venerable BJ ARIMA actually performs well. In three out of four within sample forecasts, the BJ ARIMA had smaller residuals than the CDR model. In practice, with leads over one time period (one quarter-year), the CDR has proven far more reliable because it relies on dependable forecasts on the exogenous, explanatory variables. While it is beyond the scope of this paper, which concerns combinatorics, the BJ ARIMA with the EGARCH in PROC AUTOREG could further improve this model.

In practice, the ARIMA is effective only for extremely short term forecasts. In our example, the ARIMA model uses an ar1 and an ar2 term. Hence, for forecasts projected ahead more than two quarters, the autoregressive terms will be applied to its own prior projections. The quality of the ARIMA model predictions decay rapidly. Hence, for forecasts beyond two quarters, the Dynamic Regression model will produce superior results.

Example 1: An Implementation of a Full Dynamic Regression Model with MA Terms

Constraints:

- the AR term can range from 0 to 5
- the MA term can range from 0 to 5
- two Wage and Salary explanatory variables per model
- each explanatory Wage and Salary variable can have lags from 0 to 3 time periods

Symbol Table for the Combinatoric-Generating DATA Step:

- vv represents possible Wage and Salary explanatory variables
- sym represents possible lags
- p is the order of the autoregressive term on past values of ATGR
- q is the order of the moving average on the past errors in estimation

```
data mod(keep=est);
array vv[7] $8.
("gdpdef","wsmin","ogval","wscons","wstcu","wsfire","wsserv");
array sym[4] $5. (" ", "1", "2", "3");
retain modelno 0;
do v1=1 to 7;            /* select the first wage and salary explanatory
variable                */
do v2=(v1+1) to 7;       /* select the second unique wage and salary
explanatory variable */
do s1=1 to 4;               /* select a lag for the first explanatory
variable                          */
do s2=1 to 4;               /* select a lag for the second explanatory
variable                    */
do p=0 to 5;                /* select the order of the AR term
                               */
do q=0 to 5;                /* select the order of the MA term
                               */
modelno=modelno+1;  /* assign a model identification number
                     */
est=compress("mod" || modelno)||" :  " || "estimate " || " p="||
compress(p) || " q=" || compress(q) || "  " ||
" input=(" || compress(sym[s1]) || "  " ||  compress(vv[v1]) || " " ||
compress(sym[s2]) || " " || compress(vv[v2]) || ");";
output ;
end; end; end; end; end; end;
;
run;
```

Here are three samples generated by this DATA step:

```
Sample 1:

                    1    mod1:  estimate  p=0 q=0   input=(gdpdef wsmin);
                    2    mod2:  estimate  p=0 q=1   input=(gdpdef wsmin);
                    3    mod3:  estimate  p=0 q=2   input=(gdpdef wsmin);
                    4    mod4:  estimate  p=1 q=0   input=(gdpdef wsmin);
                    5    mod5:  estimate  p=1 q=1   input=(gdpdef wsmin);
                    6    mod6:  estimate  p=1 q=2   input=(gdpdef wsmin);
                    7    mod7:  estimate  p=2 q=0   input=(gdpdef wsmin);
                    8    mod8:  estimate  p=2 q=1   input=(gdpdef wsmin);
                    9    mod9:  estimate  p=2 q=2   input=(gdpdef wsmin);
Sample 2:

                   55    mod55:  estimate  p=0 q=0   input=(1gdpdef 2 wsmin);
                   56    mod56:  estimate  p=0 q=1   input=(1gdpdef 2 wsmin);
                   57    mod57:  estimate  p=0 q=2   input=(1gdpdef 2 wsmin);
                   58    mod58:  estimate  p=1 q=0   input=(1gdpdef 2 wsmin);
                   59    mod59:  estimate  p=1 q=1   input=(1gdpdef 2 wsmin);
                   60    mod60:  estimate  p=1 q=2   input=(1gdpdef 2 wsmin);
                   61    mod61:  estimate  p=2 q=0   input=(1gdpdef 2 wsmin);
                   62    mod62:  estimate  p=2 q=1   input=(1gdpdef 2 wsmin);
                   63    mod63:  estimate  p=2 q=2   input=(1gdpdef 2 wsmin);
Sample 3:

                 3006    mod3006: estimate  p=2 q=2   input=(3 wsfire 1 wstcu);
                 3007    mod3007: estimate  p=0 q=0   input=(3 wsfire 2 wstcu);
                 3008    mod3008: estimate  p=0 q=1   input=(3 wsfire 2 wstcu);
                 3009    mod3009: estimate  p=0 q=2   input=(3 wsfire 2 wstcu);
                 3010    mod3010: estimate  p=1 q=0   input=(3 wsfire 2 wstcu);
                 3011    mod3011: estimate  p=1 q=1   input=(3 wsfire 2 wstcu);
                 3012    mod3012: estimate  p=1 q=2   input=(3 wsfire 2 wstcu);
                 3013    mod3013: estimate  p=2 q=0   input=(3 wsfire 2 wstcu);
```

** Warning: Some of the generated models use the MA term q > 0. However, some of these models may not be cointegrated.*

Example 2: Models with AR Denominator Factors

Definition of a First Order AR Denominator in a transfer function: A single parameter applied to each selected explanatory variable that "estimates the effect of an infinite distributed lag with exponentially declining weights" (SAS/ETS User's Guide, page 123).

Constraints:

- each model must have 2 explanatory variables
- the numerator (lag) must range from 0 to 3;
- the denominator factor (delta(B)) must = 1 (if used)
- the MA term must range from 0 to 2
- the AR term must range from 0 to 2

Symbol Table:

Please refer to the Symbol Table in Example One.

```
   data mod(keep=est);
   array vv[5] $8. ("gdpdef","ogval","wscons","WSRETL","wsserv");
   array sym[4] $5. (" ", "1", "2", "3","1/(1)", "2/(1)", "3/(1)");
   do v1=1 to 5;                    /* select first explanatory variable
           */
   do v2=(v1+1) to 5;               /* select non-duplicate second explanatory
   variable */
   do s1=1 to 4;                    /* select lag on first explanatory variable
           */
   do s2=1 to 4;                    /* select lag on second explanatory
   variable        */
   do p=0 to 2;                     /* select order of ar term
           */
   do q=0 to 2;                     /* select order of moving average, ma, term
           */
   est="estimate " || " p="||compress(p) || " q=" || compress(q)||"  " || "
   input=(" || compress(sym[s1])|| compress(vv[v1]) ||" " || compress(sym[s2])
   || compress(vv[v2]) ||  ") method=ml maxit=200; ";
   output ;
   end; end; end; end; end; end;
   ;
   run;
   proc print data=mod; run;

   Sample Output from data step with the AR Denominator Term included
oooo  ooo      ooo          ooo
1197  estimate p=2 q=2 input=(1/(1) wscons wsserv) method=ml  maxit=200;
1198  estimate p=0 q=  input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1199  estimate p=0 q=1 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1200  estimate p=0 q=2 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1201  estimate p=1 q=0 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1202  estimate p=1 q=1 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1203  estimate p=1 q=2 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1204  estimate p=2 q=0 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1205  estimate p=2 q=1 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
1206  estimate p=2 q=2 input=(1/(1) wscons 1/(1) wsserv) method=ml  maxit=200;
oooo  ooo      ooo          ooo
```

** Warning: Some of the generated models use the MA term q > 0. However, some of these models may not be cointegrated.*

Conclusions

Combinatoric dynamic regression is an easily implemented tool that can help the researcher find and assess all possible ARIMA models. The method demonstrated is a flexible tool for knowledge discovery. The researcher may find improved models. By using CDR, the researcher can rest assured that he has exhausted all possibilities. If improved models are not found, the researcher knows that existing models outperform any other possible models. Effectively automating the process of finding models frees the researcher to spend more time and resources on the evaluation and fine-tuning of the models found. Combinatorics may prove useful when applied to other modeling paradigms, particularly PROC MODEL and PDL.

References

SAS/ETS User's Guide, Version 6, Second Edition,
Chapter 3 ("The ARIMA Procedure") and Chapter 4 ("The AUTOREG Procedure").

Econometric Models and Econometric Forecasts, by R. Pindyck and D.L. Rubinfield, Fourth Edition, 1998. Chapters 1-11, 16-19.

Elements of Econometrics, by Jan Kmenta, 1971. Chapters 7–10.

Time Series Models, by A.C. Harvey, 1981 and 1984.

Time Series Analysis, by G.E.P. Box and G.M. Jenkins, 1976. Read the entire text; it's a classic.

Forecasting with Dynamic Regression Models, by A. Pankratz, 1983 and 1991.

Applied Statistical Forecasting, by R. L. Goodrich, 1989.

Genetic Algorithms in Search, Optimization and Machine Learning, by D. E. Goldberg, 1989.