

The Best of Both Worlds: Forecasting Using Time Series with Inputs

David A. Dickey, NC State University

ABSTRACT

This paper shows how to use regression with autocorrelated errors. It features examples using the [®]SAS procedures AUTOREG and ARIMA. Issues arising in the use of these procedures and a comparison of features of each to those of the other are presented. The emphasis is on when to use each procedure, how to understand the results, and how to use diagnostics to improve the model.

INTRODUCTION

Multiple regression is probably the most often used tool for relating a target variable of interest to a set of predictor variables. Collecting the responses into a single column \mathbf{Y} and the predictors into a matrix \mathbf{X} , the regression model is $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}$. Here $\boldsymbol{\beta}$ is a vector of unknown coefficients to be estimated, one coefficient for each predictor variable, and \mathbf{Z} is a vector of errors. The best linear unbiased estimate vector \mathbf{b} of the coefficients $\boldsymbol{\beta}$ is given by the formula $\mathbf{b}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ where \mathbf{V} is the error variance-covariance matrix. The variance-covariance matrix for the estimates is $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. Use of these formulas is referred to as using "Generalized Least Squares" or GLS. In beginning courses on regression it is often assumed for simplicity that $\mathbf{V}=\mathbf{I}\sigma^2$ where σ^2 is an assumed common variance of uncorrelated error terms and \mathbf{I} is an identity matrix. Note that the REG and GLM procedures use this simplifying assumption. This is often, but not always, reasonable. When $\mathbf{V}=\mathbf{I}\sigma^2$ the formulas simplify greatly, to $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for the estimates and $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ for their variance-covariance matrix. The term "Ordinary Least Squares" or OLS is used.

With observations taken over time, there can be autocorrelation, meaning that the matrix \mathbf{V} contains some nonzero elements off the diagonal. When this happens, or in general when there is nonzero correlation and/or unequal variances in the errors, the GLS, "Generalized Least Squares," formulas should be used. Whether it is a common variance σ^2 or a whole variance covariance matrix \mathbf{V} that applies, the elements are typically unknown. They must be estimated.

A Toeplitz matrix is one in which all elements with the same $|r-c|$ are the same where r and c are the row and column indices of the elements. In the time series case, if the data are observed at equally spaced time points (with possibly some missing values) and the \mathbf{V} matrix has the Toeplitz structure then the error terms Z_t form a stationary series. Models for Z_t are typically selected from the ARMA class of models. These have the form

$$Z_t - \alpha_1 Z_{t-1} - \alpha_2 Z_{t-2} - \dots - \alpha_p Z_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

where the q lags of ε are "moving average" terms, the p lags of Z are "autoregressive" terms. The ε sequence is assumed to be a sequence of independent mean 0, constant variance random variables, known in time series as "white noise." For forecast intervals, the procedures discussed here assume normality. Methods for estimating the unknown ARMA coefficients have been around for some time. They produce an estimated \mathbf{V} matrix to use in Estimated Generalized Least Squares or EGLS.

PROC AUTOREG: BASIC IDEA

Under some common conditions (namely that all roots of the moving average's characteristic polynomial $X^q - \theta_1 X^{q-1} - \theta_2 X^{q-2} - \dots - \theta_q$ are less than 1 in magnitude) the ARMA process is referred to as "invertible" and can be approximated arbitrarily closely by a long autoregressive process. For example, the moving average $Z_t = e_t - 0.8e_{t-1}$ has characteristic polynomial $X - 0.8$ with root 0.8. Repeated back substitution in the model gives $e_t = Z_t + 0.8e_{t-1} = Z_t + 0.8(Z_{t-1} + 0.8e_{t-2}) = Z_t + 0.8Z_{t-1} + 0.64e_{t-2} = Z_t + 0.8Z_{t-1} + 0.64(Z_{t-2} + 0.8e_{t-3}) = Z_t + .8Z_{t-1} + .64Z_{t-2} + 0.8^3Z_{t-3} + \dots$. This expresses e_t as an infinite weighted sum of current and past Z with exponentially declining weights. It shows that (1) a good estimate of e_t can be extracted from the Z series as needed for forecasting Z_{t+1} and (2) the invertible ARMA model can be approximated by a long autoregressive model with coefficients quickly approaching 0. To see this second item, note that a simple rearrangement of terms in the expression for e_t above gives the infinitely long autoregressive process as $Z_t = -0.8Z_{t-1} - 0.64Z_{t-2} - 0.8^3Z_{t-3} + \dots + e_t$. The coefficients (weights) decline exponentially fast. The largest (in magnitude) root of the moving average characteristic polynomial determines the decay rate.

Note that this condition on the roots of the moving average characteristic equation is analogous to the condition on the corresponding autoregressive characteristic equation that determines if the process is stationary. Stationarity ensures a Toeplitz \mathbf{V} matrix. The take home message here is that, as an approximation, every stationary invertible ARMA model is approximately an autoregressive process. Moving average terms serve simply to give a more parsimonious (less lag terms) model. Thus, at least to start with, an autoregressive process can be used to estimate the \mathbf{V} matrix for EGLS estimation.

Using matrices, we can see that OLS estimates are linear (linear combinations of the responses) and are unbiased, but are, perhaps, not the *best* linear unbiased estimates. The model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}$ where the elements of \mathbf{Z} form a mean 0, stationary invertible ARMA series, \mathbf{Y} , $\boldsymbol{\beta}$ and \mathbf{Z} are vectors and \mathbf{X} is a matrix. The difference between the OLS estimate \mathbf{b} and the true vector of coefficients $\boldsymbol{\beta}$ is given by the equation $\mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}) - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$. Because the expected value of \mathbf{Z} is 0, the expected value of the difference in the estimated and true $\boldsymbol{\beta}$ is 0, this being the definition of unbiased. As before, let \mathbf{V} represent the variance matrix of \mathbf{Z} which is defined as an expected value $E\{\mathbf{Z}\mathbf{Z}'\}$. Using basic facts from the theory of random vectors, the variance of \mathbf{b} is given by $E\{(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}\mathbf{Z}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ which further simplifies to the formula $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ used in PROC GLM and PROC REG in the case that \mathbf{V} is $\mathbf{I}\sigma^2$, but not otherwise. The take home message is that PROC REG and PROC GLM will give unbiased estimates of the parameters but their standard errors will be incorrect and cannot be trusted unless $\mathbf{V} = \mathbf{I}\sigma^2$. The same is true of any inference coming from them when the assumption that $\mathbf{V} = \mathbf{I}\sigma^2$ is violated. The residual vector from the OLS estimates is a reasonable estimate of the vector \mathbf{Z} of true errors and thus the OLS residual autocorrelation function helps to identify the type of ARMA model appropriate for \mathbf{Z} . If we further use the autoregressive approximation for the error series then the identification is just a matter of figuring out which lags to use.

EXAMPLE 1: (GENERATED)

Variables OURS and THEIRS simulate possible advertising expenditures in hundreds of dollars for us and for our competitor. SALES represents our company's sales, in thousands of dollars. The observations are 2 years of generated weekly data, generated from the model $\text{SALES} = 500 + 0.8 \cdot \text{OURS} - 0.5 \cdot \text{THEIRS} + Z$ where the error term Z is the stationary autoregressive order 1 time series $Z_t = 0.8Z_{t-1} + e_t$ with the standard deviation of e being 20. Figure 1 shows SALES (left Y axis) and OURS, THEIRS (right Y axis):

```

proc sgplot;
  series Y=sales X=week;
  series Y=ours X=week/y2axis;
  series Y=theirs X=week/y2axis;
run;

```

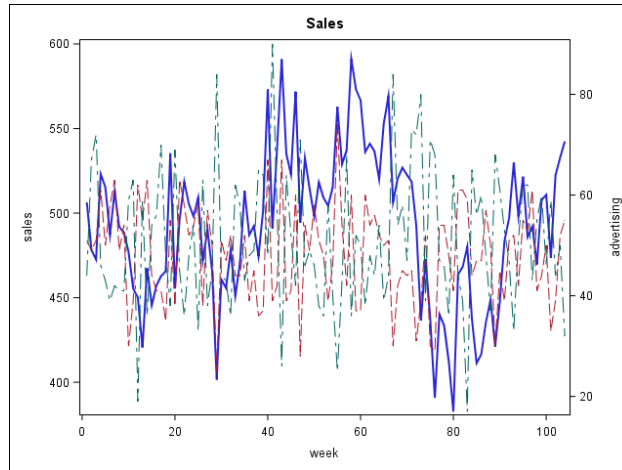


Figure 1. Sales (solid) and Advertising Expenditures (dashed).

All three series seem to have more or less constant means and variances. These are two of three characteristics of stationary series. The plot does not clearly indicate the association between the response and the predictors. Figure 2 overlays scatter plots of sales versus our advertising expenditures and versus theirs. Ordinary least squares regression lines are shown. These are just exploratory in that they ignore autocorrelation and each ignores the effect of the other advertising expenditure. They show simple correlations rather than the multivariate relationship (partial correlation). Even so, they are at least consistent with the idea that increasing our advertising increases our sales and that increasing our competitor's advertising decreases our sales.

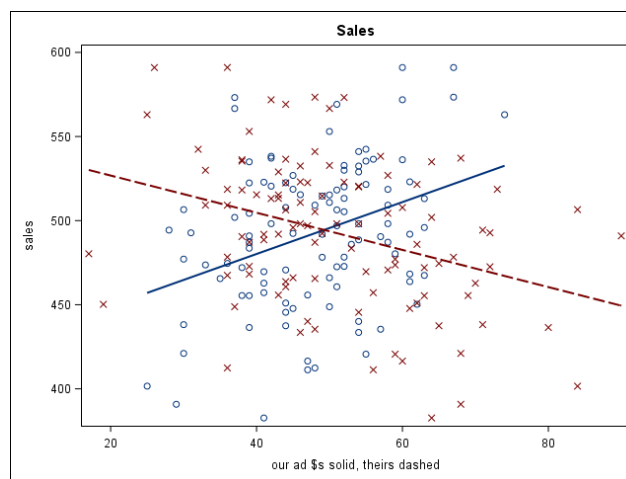


Figure 2. Our Sales Compared to Our Advertising (blue circles, solid line) and to Our Competitor's Advertising (red Xs, dashed line). Lines are Naïve Simple Linear Regression Lines.

The code follows:

```

proc sgplot noautolegend;
  title "Sales";
  reg Y=sales X=ours;
  reg Y=sales X=theirs;
  Label theirs = "our ad $s solid, theirs dashed";
run;

```

Suppose we naively run a bivariate least squares regression. The Analysis of Variance F test ($Pr > F = 0.0005$) suggests that at least one of the two advertising inputs is significant but in trying to see which of the two inputs matters, one finds p-values exceeding 0.10 for both t statistics as shown in Output 1 below. *If* these p-values could be trusted, the interpretation would be that either input could be eliminated as long as the other remains. There is, however, no justification for interpreting p-values when autocorrelation in the errors is present but not modelled. The data were generated with autocorrelated errors. The main point here is that the REG and GLM procedures are inappropriate when there is autocorrelation. Tests for autocorrelation are available.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	27536	13768	8.19	0.0005
Error	101	169811	1681.29547		
Corrected Total	103	197347			

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	
Intercept	1	483.00281	47.04252	10.27	<.0001	
ours	1	0.88580	0.59691	1.48	0.1409	
theirs	1	-0.64724	0.42609	-1.52	0.1319	

Output 1. OLS Regression of Sales on Advertising.

The AUTOREG procedure goes through three steps: (1) Run a least squares regression (matching the results above), (2) Fit an autoregressive model to the error series Z_t then (3) re-estimate the regression with EGLS using an estimated variance matrix \mathbf{V} constructed from the estimated Z_t model. In step 2, there is an option, BACKSTEP, for removing insignificant terms from the autoregressive error model:

```

proc autoreg;
  model sales = ours theirs/nlag=8 backstep;
run;

```

For these data, the BACKSTEP option correctly identifies an autoregressive order 1 model. Starting from a lag 8 model, it eliminates lag 2, then lag 8, etc. until only lag 1 remains. It estimates the error model as $Z_t - 0.857970 \cdot Z_{t-1} = e_t$. Note the way this is parameterized and that it is equivalently written $Z_t = 0.857970 Z_{t-1} + e_t$. It is a *positive autocorrelation* situation.

Backward Elimination of
Autoregressive Terms

Lag	Estimate	t Value	Pr > t
2	-0.024456	-0.18	0.8548
8	0.025465	0.25	0.8052
5	0.049289	0.38	0.7054
6	-0.049314	-0.43	0.6696
7	0.025885	0.37	0.7117
4	0.137511	1.37	0.1725
3	-0.061967	-0.79	0.4291

Output 2. Diagnosing the Z_t Series.

The MSE estimate 430.9 is reasonably close to the $20^2=400$ that was used to generate the white noise e series and the autoregressive parameter estimate 0.85797 is only about 1 standard error away from the 0.80 that was used in generating the data. The model form selected matches that used in the data generation and the model becomes

$$S_t = \beta_0 + \beta_1 O_t + \beta_2 T_t + Z_t \quad \text{where} \quad Z_t - \rho Z_{t-1} = e_t$$

Here S_t represents our Sales, O_t is Our advertising and T_t is Theirs (our competitor's advertising), each at time t . The paper on which PROC AUTOREG is based estimates the negative of ρ so here ρ is estimated as 0.85797, a positive correlation. Substituting $Z_t = S_t - (\beta_0 + \beta_1 O_t + \beta_2 T_t)$ and its lag into $Z_t - \rho Z_{t-1} = e_t$ and gathering up similar terms shows that

$$[S_t - \rho S_{t-1}] = [1 - \rho]\beta_0 + \beta_1 [O_t - \rho O_{t-1}] + \beta_2 [T_t - \rho T_{t-1}] + [Z_t - \rho Z_{t-1}]$$

so regressing (noint) $[S_t - \rho S_{t-1}]$ on $[1 - \rho]$, $[O_t - \rho O_{t-1}]$, and $[T_t - \rho T_{t-1}]$ gives a transformed regression whose parameters are precisely those that are to be estimated ($\beta_0, \beta_1, \beta_2$) and with a white noise error $e_t = Z_t - \rho Z_{t-1}$. This justifies ordinary least squares regression on these transformed [] variables. The conversion from the original variables to these transformed variables is a form of *prewhitening*. In summary, use least squares (recall the estimates are unbiased) to produce an initial Z_t series which is modelled, then use that to get improved (EGLS) estimates and more appropriate standard errors. The improved parameter estimates produce a new Z_t series and the process can be iterated. PROC AUTOREG automates this otherwise tedious process.

There are two remaining problems. One is that the first observation has no lagged value for the transformation. The other is that the autoregressive parameter ρ is unknown. Estimating ρ as described above solves the second problem. The first can be overcome as well. The AUTOREG procedure does this and reports the results of this second, prewhitened, regression as shown below. Because the parameters in the transformed regression are the same as those in the original regression, the interpretation is that every increase of 1 (hundred dollars, \$100) in our advertising is associated with a 1.0605 (thousand dollar, \$1,060.50) increase in sales and every hundred dollar increase in advertising by our competitor is associated with a \$691.00 drop in our sales. A Durbin Watson statistic near 2 suggests little evidence of remaining autocorrelation but it is important to note that the theory that underlies the distributional results of Durbin and Watson does not hold when a regression uses lagged values of the response variable as predictors. The transformed variable $[S_t - \rho S_{t-1}]$ has a lagged dependent variable in it.

Preferable checks for autocorrelation are discussed later. Output 3 shows the final parameter estimates. The initial OLS estimates are the same as in Output 1. The Durbin-Watson statistic differs little from 2, the theoretical value indicating no autocorrelation. Because it is inappropriate, a p-value for the difference, though available, was not requested. The less commonly used Durbin t and h are also available in the AUTOREG procedure.

Durbin-Watson	2.0469	Transformed Regression R-Square	0.5927
		Total R-Square	0.7812

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	478.6702	23.3229	20.52	<.0001
ours	1	1.0605	0.2432	4.36	<.0001
theirs	1	-0.6910	0.1691	-4.09	<.0001

Output 3. Partial Results from the Final Phase of the AUTOREG Procedure.

Dramatically illustrating the effect of this transformation, both parameter estimates are now very highly significant. Neither can be omitted. Reported standard errors for the slopes have changed from 0.59691 and 0.42609 to 0.2432 and 0.1691. Note that the reported OLS standard errors are not really the true standard errors that apply when OLS is used on data with correlated errors as in the first step of analysis. They come from $\mathbf{X}'\mathbf{X}^{-1}\sigma^2$, not from the correct OLS variance formula $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$. Two R-square measures are given. Recall that R-square can be computed as 1 minus the ratio of error to total sums of squares. In PROC AUTOREG the idea is to use the one step ahead error sum of squares and the usual sum of squared deviations from the mean as the total sum of squares. This gives the total R-square. It uses both the regressors and the autocorrelation in Z. The Transformed Regression (formerly labelled just "Regression") R-Square is that from the prewhitened regression. The idea here is to separate the predictive power coming from the regression from the total predictive power that also includes the autocorrelation in the Z error series.

In order to predict future values of sales, one must input future values of advertising so this would be useful for analyzing hypothetical scenarios. Our competitor is unlikely to reveal her advertising strategy. Even if forecasted values of advertising are used, the procedure has no way to incorporate their uncertainty when computing the uncertainty in forecasted sales. On the other hand, notice that even if forecasts are not of interest, one still needs to have correct standard errors for proper inference in the model and the AUTOREG procedure should be used.

To summarize and finish this section consider Table 1 of standard errors:

Table 1. Standard Errors from Different Variance Matrices.

Printed OLS from $(\mathbf{X}'\mathbf{X})^{-1}(\text{mse})$	Using σ_z^2 in OLS formula $(\mathbf{X}'\mathbf{X})^{-1}\sigma_z^2$	True OLS $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$	Printed GLS $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$	True GLS $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$
47.043	38.243	34.686	23.323	21.032
0.5969	0.4852	0.4262	0.2432	0.2406
0.4261	0.3464	0.3076	0.1691	0.1674

Items \mathbf{V} and σ_z^2 are only available because the data are generated and these true parameters known. They appear in red. The first three columns deal with OLS standard errors. The first column is what appears above in the PROC REG output. The violation of the assumption of

uncorrelated errors causes this to be incorrect. The Z_t series has a variance $\sigma_z^2 = \sigma^2 / (1 - \rho^2)$ where σ^2 is the variance of the white noise series, e_t , and $Z_t = \rho Z_{t-1} + e_t$. A theoretical version of the PROC REG standard errors can be computed as square roots of the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1} \sigma_z^2$ as shown in column 2. Column 2 shows what the estimates in column 1 are estimating. Standard errors shown in columns 1 and 2 are inappropriate because \mathbf{V} is not $\mathbf{I}\sigma^2$. As shown previously, if one uses OLS the correct standard errors are square roots of the diagonal elements of the matrix $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ using the true \mathbf{V} matrix as in column 3. The point here is that ordinary least squares procedures like the REG and GLM procedures do not give valid standard errors, t tests, or p-values when \mathbf{V} is not $\mathbf{I}\sigma^2$.

Columns 4 and 5 give estimated (from the printout) and theoretical (available because the data are generated) standard errors for the GLS estimates of the parameters. An interesting comparison is between columns 3 and 5. This comparison illustrates the theoretical result that the true standard errors from GLS are smaller than those (correct ones) from OLS. The *true* GLS variances of parameter estimates will never exceed those for OLS. This can be confusing because the *printed* standard errors from REG or GLM are sometimes smaller than those from GLS which would *seem* to suggest that the OLS estimates are superior. That conclusion is unjustified. The printed OLS standard errors are not estimating the true standard deviations of the estimated regression parameters. Comparing printed standard errors from REG or GLM to those from AUTOREG to see which is best is *not* justified.

EXAMPLE 2: DAILY RIVER FLOWS, NEUSE RIVER IN NC

The Neuse River flows through North Carolina and into the Atlantic Ocean, passing first through Goldsboro NC then Kinston NC on its way. As is typical with series like this that cannot be negative but have values near 0 along with high ones, there is more variation in the high flow rates than the low ones, suggesting heterogeneous variances and thereby violating an important model assumption. Sometimes analyzing natural logarithms of such data will alleviate the problem. An added benefit of analysis on the logarithmic scale is that predicted values converted to the original scale will never be negative regardless of the sign of the logarithm.

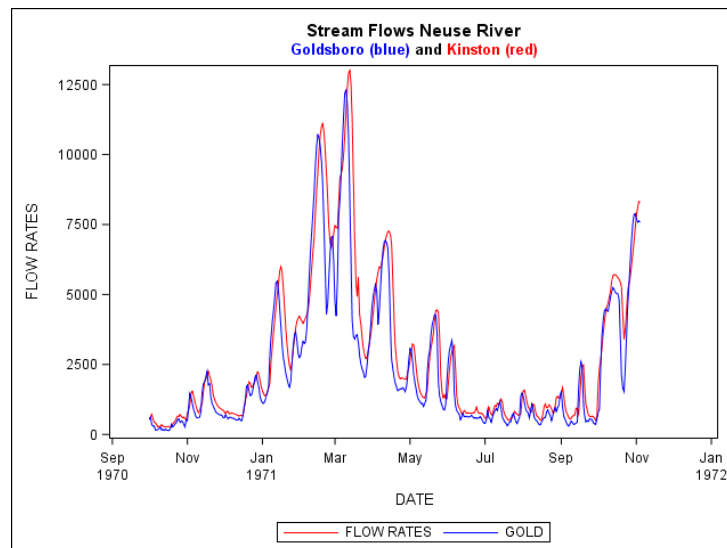


Figure 3. Neuse River Flow Rates at Goldsboro and Kinston NC, Original Data.

Streamflows were taken on 400 consecutive days at Goldsboro (the lower blue curve) and downstream at Kinston (the upper red curve). The plot suggests analysis of the data on the logarithmic scale. The corresponding (base e) log scale data are plotted in Figure 4. In Figures 3

and 4, the downstream location can be identified by noticing that its peaks seem to occur one day later than those of the upstream location and the flows are also larger, as would be expected from a downstream location.

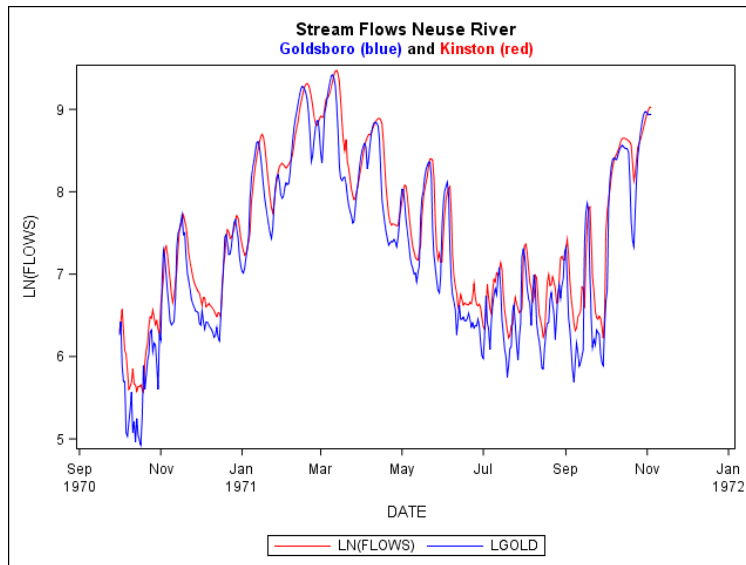


Figure 4. Log Transformed Flows at Goldsboro (lower curve) and Kinston NC.

In Figure 4, an almost sinusoidal trend appears. This suggests that the sine and the cosine of $2\pi t/365.25$ might be useful predictor variables. The AUTOREG procedure code for the log transformed Goldsboro flows (Lgold) is given below. A similar code for Kinston, Lkins, was run as well:

```
proc autoreg data=river;
  model Lgold = sine cos /nlag=5 backstep ;
  output out=outgold
  predicted=plgold pm=pmlgold residual=rlgold rm=rmlgold;
```

Output 4 shows that a lag 2 autoregressive model is chosen.

Estimates of Autoregressive Parameters

Lag	Coefficient	Standard Error	t Value
1	-1.305266	0.046737	-27.93
2	0.370404	0.046737	7.93

Output 4. Autoregressive Error Model is $Z_t - 1.3053Z_{t-1} + 0.3704 = e_t$

The autoregressive (AR) characteristic polynomial is $X^2 - 1.305266X + 0.370404$ (recall the AUTOREG parameterization $Z_t + \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} = e_t$). Figure 5 shows this polynomial for log transformed data from both stations.

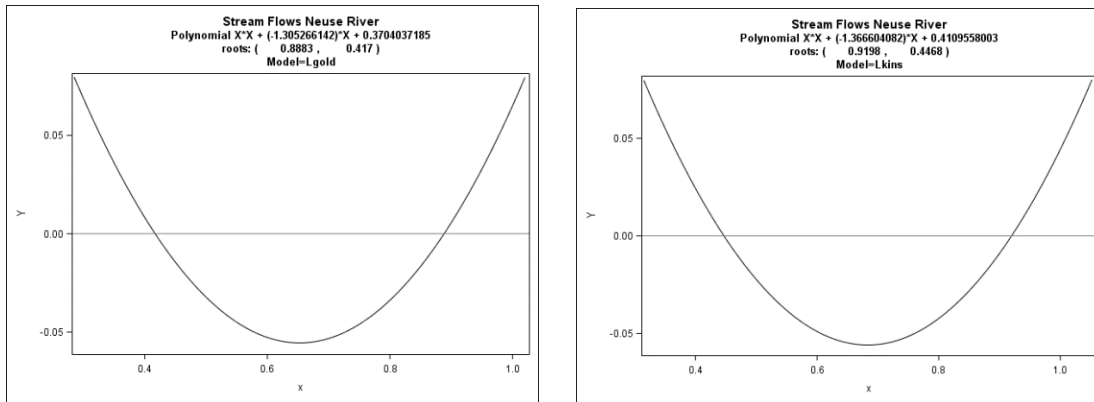


Figure 5. AR Characteristic Polynomials, Goldsboro (left) and Kinston (right).

The largest root (rightmost crossing of 0) of the Goldsboro polynomial is 0.8883 and that of Kinston is 0.9198. Both are somewhat close to 1. Such a root causes the next forecasted residual to be near the current residual and suggests that the estimated Z variance will be much larger than that of the white noise e variance. This phenomenon manifests itself in Output 5 where the Total R-Square, 0.9636, which uses both the sine-cosine function *and* the error model for Z, is much larger than the Model R-Square, 0.0499, that uses only the sine-cosine function. The interpretation is that most of the one step ahead forecasting performance is due to the autoregressive model with a root near 1 (that is, the momentum in the residuals). The trigonometric inputs seem to provide little additional explanatory power. On the other hand it is common knowledge that temperatures, rainfall, and streamflows have a seasonal pattern that is somewhat sinusoidal in nature. The trigonometric functions will thus remain in the model.

Durbin-Watson	1.8831	Transformed Regression R-Square	0.0499
		Total R-Square	0.9636

Output 5. More AUTOREG Output.

The DWPROB option gives the probability that there is autocorrelation. It is justified if the model does not involve lagged dependent variables. Lagged dependent variables enter the model through the use of the prewhitening transformation, so the p-values are only rigorously validated in the initial OLS regression. The DWPROB option is thus not used here.

The final sinusoidal model from PROC AUTOREG is given in this output:

Parameter Estimates						
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t	
Intercept	1	7.3563	0.1533	48.00	<.0001	
SINE	1	0.8113	0.2060	3.94	<.0001	
COS	1	0.5598	0.2222	2.52	0.0121	

Output 6. Final Sinusoid Parameter Estimates.

The output variables that have M (for mean) in their names, such as the "PM=" or "RM=" variable, predict only from the regressors and do not take advantage of the momentum in the residuals. The PM prediction is $7.3563 + 0.8113 * \text{SINE} + 0.5598 * \text{COS}$ and the RM residual is the observed value minus this PM prediction. Because the sine and cosine variables are deterministic they can easily be extended into the future and thus the associated PM predictions can be computed. Figure 6 shows these PM trigonometric predictor functions, both throughout

the historical data and into the future several months. Because the data end on a relatively high value and the Z_t model has a root near 1, it takes a couple of months for the forecasts to become visually indistinguishable from the deterministic sinusoidal functions.

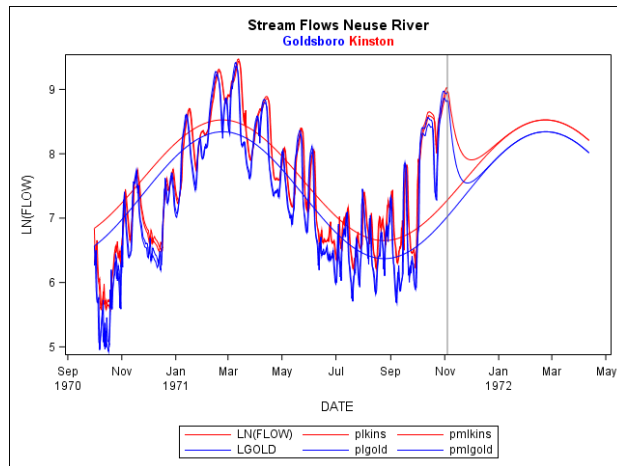


Figure 6. Log of Flow: Historical Data and Forecasts, With and Without Accounting for Autocorrelation.

Figure 7 shows the RM residuals (estimates of Z_t) for both stations as solid lines and the R residuals (estimates of e_t), i.e. the deviations from forecasts that use the regressors and the error model, as circles. This shows the rather large reduction in forecast error variance due to the use of the error model to forecast the next Z . The circles are much more tightly clustered around 0 than are the points on the series plots. This further illustrates the difference between the total and model R-square values.

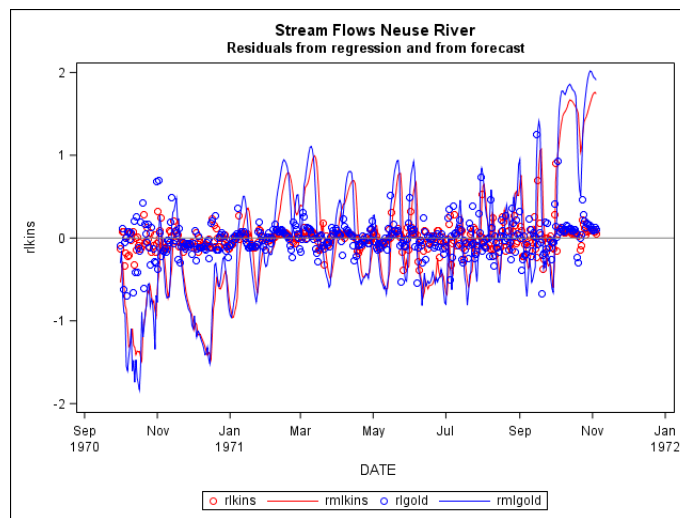


Figure 7. The Z_t (lines, r_m) and e_t (circles, r) Residuals over Time.

Large negative residuals r_m on the left and large positive ones on the right suggest possibly adding a trend. This improves the fit but suggests a long term large linear increase in flow that is nonsensical. Similarly a unit root model could be fit but the implied unbounded variance for such models is also an untenable assumption for a river.

DISCUSSION AND ALTERNATE MODEL FOR KINSTON:

The river example illustrates several things about analysis in general and time series in particular. River flows, rainfall, and other seasonal series typically have a yearly cycle of some kind. Periodic predictors, like the sine and cosine pair used here, are well motivated. While a linear trend makes the fit look better in this example, it would be hard to believe that river flows will follow a linear trend over time. Common sense must come into play. It is possible to find a model that uses Goldsboro flows to predict the downstream flow at Kinston. From the graphs it is obvious that using Goldsboro to predict Kinston will give excellent one step ahead forecasts in the historical data and one step into the future but how do we forecast further with such a model? This is the problem with using stochastic inputs rather than deterministic inputs like the sine and cosine whose future values are known perfectly. On the logarithmic scale two Goldsboro lags supply most of the predictability for Kinston. The problem in forecasting more than 1 step ahead is that we do not know what the required Goldsboro inputs will be for insertion into the predicting equation. One possible solution is to use the Goldsboro predictions from the sine and cosine analysis that produced Figure 6 whenever future Goldsboro flows are needed. PROC AUTOREG has no way of knowing the uncertainty induced by using these predicted flows rather than actual future Goldsboro flows so any prediction intervals computed will be too narrow. The following code fits the model under discussion:

```
proc autoreg data=laggold;  
  model lkins = lgold_1 lgold_2 /nlag=15 backstep;  
  output out=out2 pred=p pm=pm residualm=r residual=r; run;
```

where the laggold data set has variables lgold_1 and lgold_2 that hold the lag 1 and 2 log(Goldsboro flow) values in the historical data and predictions as just described for the future. The resulting predictions and actual values are plotted, in Figure 8, as in Figure 6.

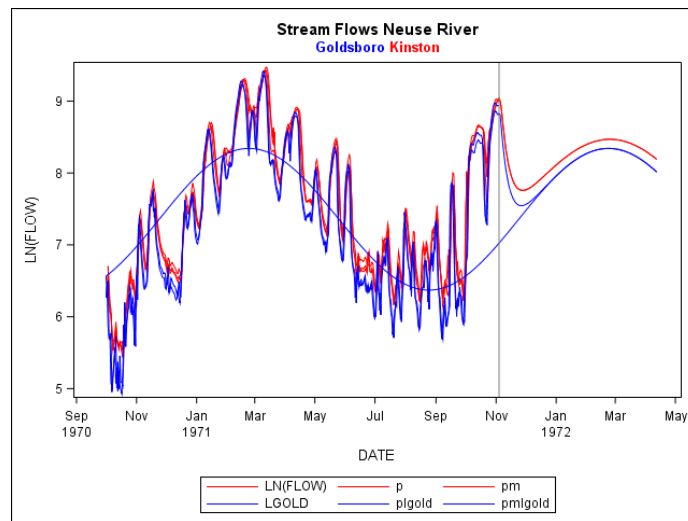


Figure 8. Kinston Predictions using Lagged Goldsboro Flows.

The predicted mean function for Kinston no longer appears as a sine wave in the historical data. That is because the one step ahead Kinston predictions are computed from observed lagged Goldsboro values, not sine and cosine values. They follow the observed Kinston values much more closely. The sinusoidal Goldsboro predictions are shown throughout just as in Figure 6 but only the future sinusoidal values for Goldsboro are used in forecasting.

Important point: Excellent one step ahead forecast performance in the historical data does not imply excellence in the future forecasts more than one or two steps out. In the opinion of this author, there is not much improvement here over the deterministic sine and cosine approach previously shown. One difference to note is that in Figure 8, the future Kinston forecasts are closer to the Goldsboro forecasts than they are in Figure 6. This is due to the use of the Goldsboro forecasts (that approach the sinusoidal curve) in place of actual future Goldsboro values and the dependence of the Kinston forecasts on these forecasted future values. In fact the total R-square for this model exceeds 99%. It seems that when 2 actual previous values of Goldsboro flow are available an excellent one step ahead Kinston forecast results. Clearly a 99% R-square in the historical data and does not suggest excellent forecasts far into the future.

Example 3: Energy Demand, NC State University

This is an excellent example of PROC AUTOREG working well and illustrates some additional features not yet discussed. The data are real as were the river flows. The data are historical daily demands for energy at NC State University during the 1979-80 school year. A few years earlier an oil embargo had caused gasoline shortages and concern about energy usage in general. Now a second shortage was in progress. NC State University posted a sign showing the previous day's usage at the campus entrance each day. Upon request, the Facilities Division sent the usage numbers on paper. These numbers and data from the academic calendar were entered. The academic year features (1) vacation days and weekends, (2) work days, and (3) class days. Class days are a subset of work days. A sine and cosine of period 1 year capture seasonal effects as in the river data. A third data source supplied daily temperatures. The type of day and values of the sine and cosine are perfectly known into the future but temperature is a stochastic process and future values are not perfectly known. If the sine and cosine pick up enough seasonality to substitute for temperature, problems with stochastic inputs would be avoided.

Figure 9 presents a graph of the energy usage data against time (left) and temperature (right). Vacation days and weekends are plotted in blue (lowest curve and scatter), non-class work days in green (middle group) and class days in red (top). Overlaid on the observations are exploratory least squares fits of parallel sinusoids (left) and quadratics in temperature (right). Each predictor appears effective alone, but is either still important in the presence of the other?

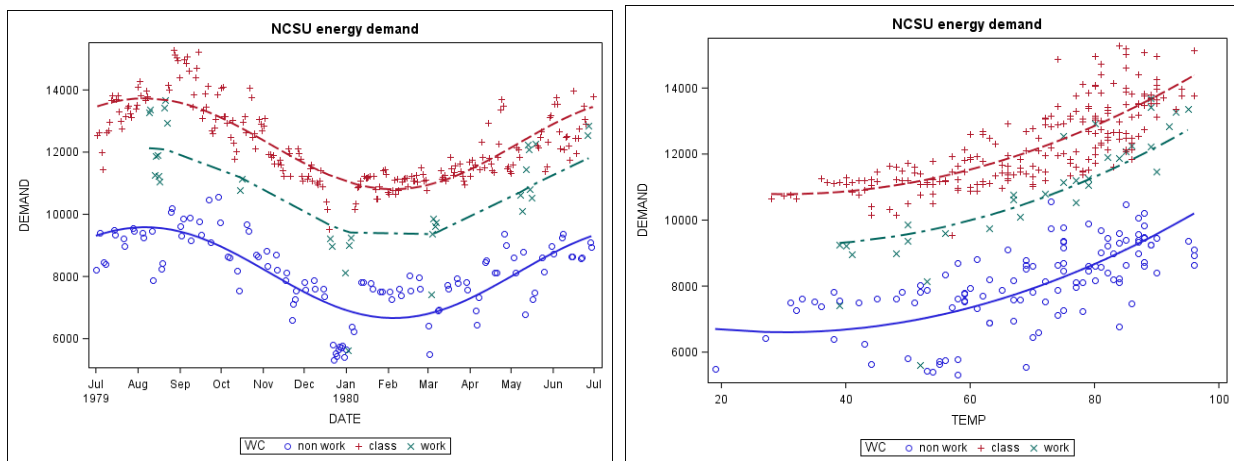


Figure 9. Energy Usage Versus Time (left) and Temperature (right).

The following code gives an initial analysis showing all sources discussed thus far are significant:

```

proc autoreg data=energy;
  model demand = temp tempsq class work s c
    /nlag=15 backstep dwprob;
  output out=out3
  predicted = p predictedm=pm
  residual=r residualm=rm;
run;

```

Recall the steps in the fitting process. After fitting an initial OLS model, the residuals are diagnosed and an estimated GLS fit is performed based on the diagnosed correlation structure. Below is the selected autoregressive model structure

Estimates of Autoregressive Parameters			
Lag	Coefficient	Standard Error	t Value
1	-0.559658	0.043993	-12.72
5	-0.117824	0.045998	-2.56
7	-0.220105	0.053999	-4.08
8	0.188009	0.059577	3.16
9	-0.108031	0.051219	-2.11
12	0.110785	0.046068	2.40
14	-0.094713	0.045942	-2.06

Output 7. Energy Data Autocorrelation Diagnosis.

Some of the lags are intuitive like lags 1, 7, and 14. Other lags are not so intuitive and there are quite a few of them. Some of the autocorrelation might be due to a day of the week effect that was not modelled. Next we see the parameter estimates from the GLS fit in which standard errors and t tests are valid. It seems that all of the effects in the model are highly significant.

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	6076	296.5261	20.49	<.0001
TEMP	1	28.1581	3.6773	7.66	<.0001
TEMPSQ	1	0.6592	0.1194	5.52	<.0001
CLASS	1	1159	117.4507	9.87	<.0001
WORK	1	2769	122.5721	22.59	<.0001
S	1	-764.0316	186.0912	-4.11	<.0001
C	1	-520.8604	188.2783	-2.77	0.0060

Output 8. Final Estimates and Tests.

In this table, S and C are the sine and cosine. Temperature and its square appear and there are two dummy variables. CLASS indicates the effect of a day in which classes are in session. WORK indicates a work day. Since the "weekend and vacation" dummy variable is omitted, the two dummy variable coefficients are deviations from that effect. They describe the shifts between the parallel curves in Figure 9. On days in which classes are in session, both the WORK and CLASS dummy variables are 1 so that the predictions on those days are $1159 + 2769 = 3928$ units above the weekend predictions. The Total R-square for this model is just over 95%, an impressive number but recall that this is obtained for known temperatures whose future values are unknown. Having asked for residuals (rm) from the model that uses input variables only, we next plot them in PROC SGPLOT with the NEEDLE statement and a reference line at $rm=0$. Unlike the r residuals, the rm residuals are autocorrelated.

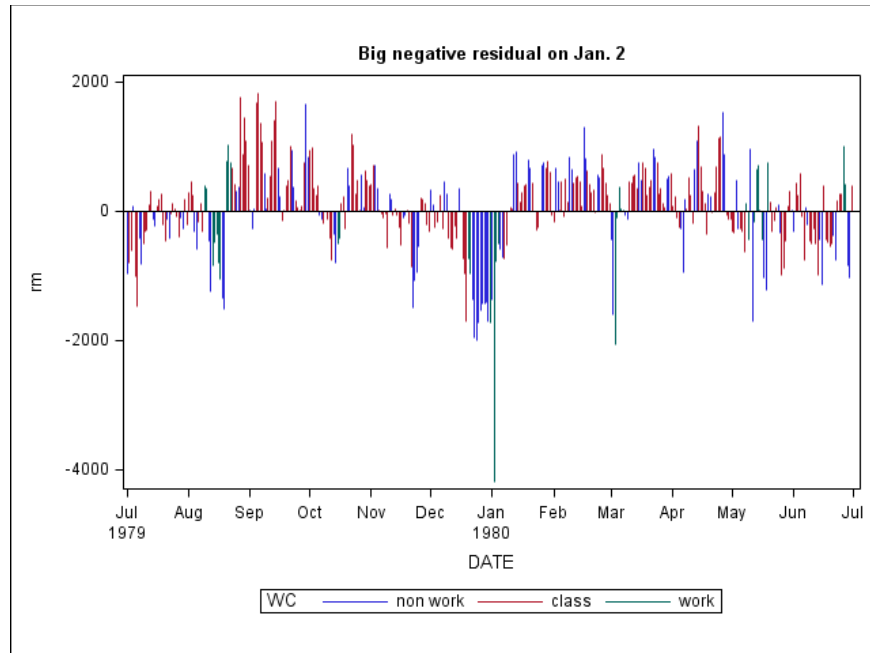


Figure 10. Autocorrelated Residuals $rm (Z_t)$.

A striking feature is the very large negative residual on January 2. The legend indicates that this was a work day but not a class day. Looking back at Figure 9 carefully we see the corresponding green X within the cluster of blue (non workday) points near January 2 in the left plot and at about 70 degrees in the right plot. It seems January 2 was a work day but its effect on energy was like another vacation day. There were no classes on January 2 so many employees may have chosen to use a vacation day.

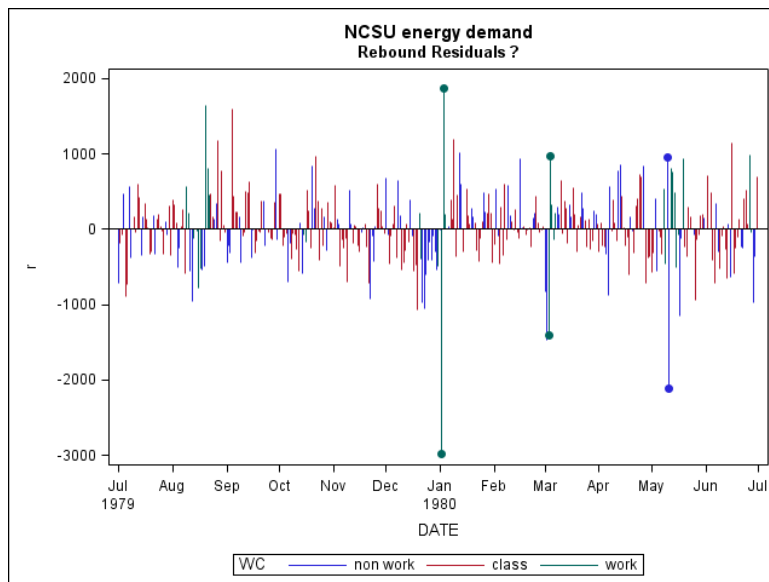


Figure 11. White Noise Residuals $r (e_t)$ from Predictions that use Inputs and Autocorrelation.

In Figure 11, the forecast residuals, those that result from using both the inputs and the autoregressive error structure, are shown. Strangely, the unusually low value on January 2 is

followed by a large residual on January 3. Two other outliers identified by PROC ARIMA are shown (as thick lines with dots on the ends) each with a similar “rebound outlier” following it. Why is this happening? The answer lies in the positive lag 1 autocorrelation. On January 2 there is an exceptionally low residual so the model predicts that the next observation will be far below what is predicted by the X variables alone. When it turns out that the next observation is closer to what is predicted by the inputs, a positive “rebound” residual is produced. It is not that the observation is too high but rather that the prediction is too low.

Because the autoregressive lag structure is somewhat complex and includes some unexpected lags, the next step is to move to the more sophisticated ARIMA procedure. When inputs are present, one must include them in the CROSSCOR list as well as in the INPUT= option of the ESTIMATE statement. Perhaps with the added possibility of moving average terms a more explainable lag structure will arise. ARIMA also has an outlier detection feature. The data set has 366 observations. We might have anticipated a January 2 effect but for “discovering” other outliers it is prudent to account for the multiple tests that we are doing. A conservative way to do this is to insist on a p-value less than 0.05/365 as suggested by the Bonferroni correction for declaring a point an outlier. Each of 365 points is being tested, hence the denominator 365.

The following code shows the PROC ARIMA analysis. It includes the search for outliers:

```
proc arima data=energy;
  identify var=demand crosscor=(temp tempsq class work s c) noprint;
  estimate input = (temp tempsq class work s c);
  estimate input = (temp tempsq class work s c) p=(1)(7) q=(14) ml;
  forecast lead=0 out=outarima id=date interval=day;
    * 0.05/365 = .0001369863 (bonferroni) *;
  outlier type=additive alpha=.0001369863 id=date;
run;
```

The first ESTIMATE statement fits a least squares regression, much like what PROC AUTOREG does and it produces diagnostic plots from which an ARMA model diagnosis can be made. Doing so, along with some trial and error, produces a more intuitive structure: a multiplicative autoregressive structure that produces lag 1, 7, and 8 effects, and an almost significant (at the traditional 0.05 level) moving average coefficient at lag 14. Here are the resulting parameter estimates:

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag Variable
MU	6256.0	298.89104	20.93	<.0001	0 DEMAND
MA1,1	-0.10191	0.05489	-1.86	0.0634	14 DEMAND
AR1,1	0.67921	0.04048	16.78	<.0001	1 DEMAND
AR2,1	0.19973	0.05427	3.68	0.0002	7 DEMAND
NUM1	26.07617	3.85393	6.77	<.0001	0 TEMP
NUM2	0.63087	0.12242	5.15	<.0001	0 TEMPSQ
NUM3	960.27699	123.88594	7.75	<.0001	0 CLASS
NUM4	2914.3	126.20869	23.09	<.0001	0 WORK
NUM5	-771.98511	162.30684	-4.76	<.0001	0 S
NUM6	-573.09106	171.88681	-3.33	0.0009	0 C

Output 9. Analysis using the ARIMA Procedure.

An advantage of the ARIMA procedure is its sophisticated collection of diagnostics along with its ability to use moving average terms in the error model and to handle differencing when needed. As mentioned, diagnostics superior to the Durbin Watson test are available. They are based on the autocorrelation in the residuals. The ARIMA and AUTOREG procedures use residual correlation to produce the forecasts. If there remains correlation in the residuals why was it not used? It must mean that the chosen model is not rich enough, that is, a good model should result in 0 residual autocorrelation. Box and Pierce (1970) and Ljung and Box (1978) developed the Chi Square test that is used in PROC ARIMA. Box's student Pierce used the asymptotic normality of autocorrelations to show that the sum of k squared autocorrelations, multiplied by n, has approximately a Chi-square distribution in large samples under the null hypothesis of no residual correlation. It is a lack of fit test. Under the null hypothesis the chosen autocorrelation structure is sufficient while under the alternative the fit is not good enough, a lack of fit. As with any lack of fit test, a good model should result in a p-value *larger* than 0.05.

Following Pierce's work, another Box student, Greta Ljung, showed that multiplying the j^{th} term in Pierce's sum by $(n+2)/(n-j)$ provided better finite sample performance and it is her version of the test that appears in PROC ARIMA. The test is sometimes referred to as the Ljung-Box Q statistic. Neither Pierce nor Ljung suggested how many squared correlations to sum. SAS procedures show the statistics and approximate p-values for several cumulative sums. Output 10 shows the results for the energy data.

From Output 9, all parameter estimates are significant at the 0.05 level with one exception having $p=0.0634$. The lack of fit tests look reasonable even out to lag 48 (48 normalized squared residual autocorrelations summed)

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	4.19	3	0.2415	-0.044	0.024	-0.023	0.019	0.060	0.066
12	11.30	9	0.2555	-0.004	-0.012	0.121	0.044	-0.028	-0.038
18	13.91	15	0.5327	-0.051	0.009	-0.023	0.044	-0.007	-0.040
24	15.97	21	0.7711	-0.021	-0.024	0.012	-0.054	0.020	-0.028
30	24.44	27	0.6061	0.008	0.055	-0.103	0.080	-0.003	0.034
36	35.38	33	0.3564	-0.021	-0.072	0.066	-0.005	0.122	-0.044
42	40.09	39	0.4215	0.005	-0.007	0.090	0.012	0.001	0.055
48	43.59	45	0.5319	-0.042	0.044	-0.027	-0.048	-0.023	-0.031

Output 10. Lack of Fit Results.

All of these results support the model but there is still the problem of unaccounted for outliers. The outlier statement produced these three results:

Outlier Details						
Obs	Time ID	Type	Estimate	Chi-Square	Approx Prob> ChiSq	
186	02-JAN-1980	Additive	-3225.4	97.10	<.0001	
315	10-MAY-1980	Additive	1843.1	34.03	<.0001	
247	03-MAR-1980	Additive	-1509.8	22.83	<.0001	

Output 11. Outliers Using the ARIMA Procedure (Bonferroni adjustment).

The January 2 result was expected. The positive outlier in May occurred on a Saturday and another look at the academic calendar revealed that this was graduation day. The negative

outlier in March remained a mystery until an internet search of local newspaper headlines revealed that one of the largest snowstorms in NC history occurred on the Sunday before this Monday class day. An unsubstantiated but quite likely guess is that classes were called off that day. These outliers can be explained which suggests including a dummy variable for each. Because the dummy variable will devote a parameter to capturing each of these effects (Jan 2, graduation, and the storm) the three residuals should now be smaller in magnitude, possibly eliminating the rebound effect. That is, it is not advisable to put dummy variables in the model for those large rebound effects. They are an artifact of ignoring the three outliers. Here are the parameter estimates for the final model with January 2 facetiously called "hangover day."

The ARIMA Procedure

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MU	6127.4	259.43918	23.62	<.0001	0	DEMAND	0
MA1,1	-0.25704	0.05444	-4.72	<.0001	7	DEMAND	0
MA1,2	-0.10821	0.05420	-2.00	0.0459	14	DEMAND	0
AR1,1	0.76271	0.03535	21.57	<.0001	1	DEMAND	0
NUM1	27.89783	3.15904	8.83	<.0001	0	TEMP	0
NUM2	0.54698	0.10056	5.44	<.0001	0	TEMPSQ	0
NUM3	626.08113	104.48069	5.99	<.0001	0	CLASS	0
NUM4	3258.1	105.73971	30.81	<.0001	0	WORK	0
NUM5	-757.90108	181.28967	-4.18	<.0001	0	S	0
NUM6	-506.31892	184.50221	-2.74	0.0061	0	C	0
NUM7	-3473.8	334.16645	-10.40	<.0001	0	hangover	0
NUM8	2007.1	331.77424	6.05	<.0001	0	graduation	0
NUM9	-1702.8	333.79141	-5.10	<.0001	0	storm	0

Output 12. Incorporating Three Special Cases (outliers).

The January 2 effect, a reduction of 3474 in energy usage, is near the difference 3258 in going from a vacation day to a work day, that is, the expectation that January 2 would be taken as a vacation day seems consistent with these results. Note too that the previously near significant MA1,2 term at lag 14 is now significant with $p=0.0459$. Have the good diagnostics been retained with this change? The table of Ljung-Box Q statistics below shows large (supportive) p-values like those in the previous model that ignored the outliers, except in the last two rows where the previous model had larger (more supportive) p-values. It seems that a large residual correlation or two (like the 0.122 at lag 42) have caused the Q statistic to become large. Inserting a moving average term or dummy variable at a large lag like 42 seems, to this author, unjustified and will not be pursued.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	2.74	3	0.4342	0.035	-0.034	-0.046	-0.024	0.047	-0.005
12	9.16	9	0.4225	-0.001	-0.006	0.120	0.021	-0.046	0.010
18	16.93	15	0.3229	-0.105	0.019	-0.047	0.080	-0.002	-0.011
24	21.41	21	0.4340	-0.034	-0.040	0.056	-0.072	-0.018	0.002
30	31.71	27	0.2431	0.031	-0.004	-0.077	0.123	-0.045	0.043
36	42.34	33	0.1279	-0.036	-0.092	0.034	0.049	0.109	-0.031
42	57.33	39	0.0293	0.004	-0.037	0.100	0.022	-0.096	0.122
48	61.21	45	0.0541	0.000	0.022	0.020	-0.060	-0.046	-0.051

Output 13. Lack of Fit (residual check) for New Model.

EXAMPLE OF ARCH MODELS: DOW JONES

For some stock series the closing prices Y_t are used to compute an overnight return approximation $\ln(Y_t/Y_{t-1})$ which is approximately the daily proportional change in the closing price. In a model with only a mean as input, such as what is about to be presented for the Dow Jones Industrials Average, the square of each residual can be thought of as an estimate of the local error variance. Because forecasts coming from ARIMA models are linear combinations of past data points, the application of such models to these squared return residuals can provide smoothed local variance estimates that depend mostly on their recent predecessors and relatively little on squared residuals from the distant past. The data used here are historical Dow Jones returns in the first part of the 20th century. The data are of historical interest in that there were periods of great stock market volatility as well as calmer periods – exactly the type of data for which ARCH (AutoRegressive Conditionally Heteroscedastic) models and their relatives (GARCH, IGARCH, EGARCH, etc.) are meant. Figure 12 on the left shows the Dow Jones returns and the right panel shows prediction bands in blue.

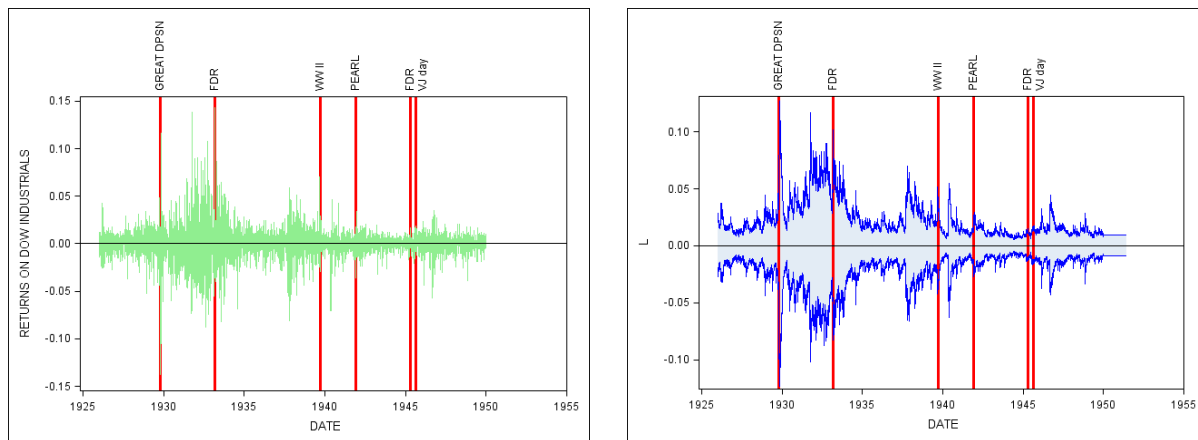


Figure 12. Dow Jones Returns (left) and IGARCH Prediction Bands (right).

The red vertical reference lines are, left to right, the Great Depression, FDR enters office, WW II starts, Pearl Harbor, FDR leaves office, and VJ day. The ARCH model, developed by Robert Engle, and its relatives use a likelihood in which the variance h_t at time t follows an ARIMA model. The following code produces the analysis by fitting a 2 lag autoregressive process with an integrated GARCH(2,1,1), or IGARCH, model. The middle argument, 1, indicates a unit root model (the INTEG option in PROC AUTOREG) for the local variance h_t . As usual, a unit root model produces forecasted variances that are very close to their predecessors making the ARCH model very responsive to changes. In unit root models, an intercept in the differences gives a nonzero slope in the original levels, which are variances here, so it is prudent to use the NOINT option unless it is clear that the variance is increasing or decreasing linearly:

```
proc autoreg data=dowjones;
  model ddown = / nlag=2 garch=(p=2,q=1,type=integ,noint);
  output out=out2 predicted=f lcli=l ucli=u;
run;
```

Graphing the prediction limits L and U separately from the data, in Figure 12, avoids the excessive overlap that would arise from overlaying the data. Notice that the prediction limits into the future appear consistent with the most recent interval widths. These data as well as the

NCSU energy demand and Neuse River data are discussed in more detail in [SAS for Forecasting Time Series 3rd ed.](#) by Brocklebank, Dickey, and Choi (2018).

CONCLUSION

The AUTOREG and ARIMA procedures provide a time series practitioner with the ability to use both predictor variables and autocorrelation to produce forecasts. The effect of autocorrelation is important in the short run but forecasts far out in the future are determined almost exclusively by the inputs as in Figures 6 and 8, the river data. Deterministic predictors have the obvious advantage of known future values. AUTOREG is easy to use and serves as a good initial modelling tool even if ARIMA is later used as in the campus energy data. Advantages of ARIMA include the inclusion of moving average terms in the autocorrelation model, differencing (not illustrated here) and the detection of outliers.

REFERENCES

Box, G. E. P.; Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models". Journal of the American Statistical Association. 65 (332): 1509–1526.

Brocklebank, J. C.; Dickey, D. A.; and Choi, B. (2018). SAS for Forecasting Time Series 3rd ed. (SAS Institute).

Ljung, G.M. and Box, G.E.P. (1978). "On a Measure of a Lack of Fit in Time Series Models". Biometrika. 65 (2): 297–303.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David A. Dickey (Emeritus Professor of Statistics, NC State University)
dickey@stat.ncsu.edu