Regression with Time Series Errors

David A. Dickey, North Carolina State University

Abstract:

The basic assumptions of regression are reviewed. Graphical and statistical methods for checking the assumptions are presented using a sales example. Departures from independence in time series data are emphasized and illustrated in the example. Several products from SAS™ Institute for analyzing regressions with time series errors are illustrated. The importance of the stochastic properties of the model input variables is emphasized. Forecasts from several models for the example data are compared.

1. Introduction:

Regression is a tool that allows one to model the relationship between a response variable Y, which might be a mail order company's sales, and some explanatory variables usually denoted X_j where X_1 might be the cost of one item from the company, X_2 the cost of a similar item from a competitor company and X_3 the number of phone calls coming in to the company's switchboard. A typical regression model for this situation is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$$

where the regression coefficients, β, are unknown.

You would like to estimate these βs, for if you could, you would then have an equation for predicting a future Y from the associated Xs. Notice that even if the regression coefficients were known, such a prediction would require knowledge of future X values. For example, if t represents time and $X = t$, then part of our model consists of a simple linear time trend and there will be no surprises when we try to extend the time sequence $1, 2, 3, \ldots, n$ into the future. On the other hand if X_t is the number of incoming phone calls at time t then forecasting to time $n + 1$ would require that some value be inserted for X_{n+1} and this value will itself likely be a forecast. These two examples represent *deterministic* and *stochastic* explanatory variables, respectively.

The nature of the X variables will affect the forecast accuracy - obviously a person forecasting with a known future X is better off than one who must estimate that future X. Thus a problem we will need to deal with, if we want to put some sort of error bounds on our forecasts, is the incorporation of our level of uncertainty about the future X values.

The usual way of estimating the βs is the method used in PROC GLM and PROC REG. The method is referred to as *ordinary least squares* in that it finds estimates b_j of the β_j parameters that minimize the *error sum of squares* SSE $= \sum_{t=1}^{n}(Y_t - (b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3))^2$. This SSE is a function of the estimates, b_j, and much of the subject of calculus is concerned with finding values of arguments, like these b$_j$, that minimize a function, SSE in our case. Thus we have mathematical tools which are relatively easy to implement on the computer that allow us to find the minimizing values. This is what PROC REG and PROC GLM are set up to do. Furthermore, statistical theory allows us to compute measures of uncertainty called *standard errors* for these b_j estimates and the resulting forecasts if certain conditions are satisfied. Note the expression "if certain conditions are satisfied." It is this with which we are concerned here.

In this paper we review these "certain conditions," indicate why they might be violated when data are taken over time, present methods for checking these conditions, and finally represent corrections that can be applied if the conditions are violated. The corrections that we speak of are implemented in SAS Institute's PROC AUTOREG.

Throughout the paper we will use an artificial example in which X_t represents the number of phone calls in week t to a mail order company and Y_t is the number of shipments for that week. Figure 1 shows the data over a 3 year period. We are interested in estimating the company's growth, estimating the number of shipments generated per phone call, and forecasting phone calls and sales two weeks into the future.

2. Checking the usual assumptions.

Our model is $Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + e_t$. We assume

(A) Normality:
The errors all come from normal
distributions

(B) Homogeneity:
These normal distributions all
have mean 0 and
the same variance, σ^2

(C) Independence:
The correlation between e_i and e_j is 0
(for i not equal to j)

We can check the normality assumption by drawing histograms and normal probability plots of the residuals. In figures 2-4, we see a histogram of the residuals, a hanging histogram in which each bar becomes a line segment at the former bar midpoint, this line being hung from the normal curve rather than rising from the horizontal axis, and a plot of the residuals against their normal scores. These are very easy to produce using the following code:

```
proc capability graphics;
histogram r /normal hanging vref=0;
histogram r / normal;
qqplot r / normal (mu=est sigma=est);
```

The histograms look reasonably normal and the quantile-quantile plot reasonably straight. PROC CAPABILITY also presents tests of the normality hypothesis but the theory behind these assumes independence, an assumption we have yet to check.

Not shown is a simple plot of residuals against predicted values. Because this looks uniform (as opposed to megaphone shaped) this check on the homogeneous variance assumption does not give us cause for concern.

The regression and subsequent calculation of residuals was accomplished with this code:

```
proc reg; model y = t x/dw;
proc reg; model y = t/dw;
```

where Y is sales, X phone calls, and t week number. The previous residual analysis was from the first regression. The advantage of the second regression is that only future values of t would be needed for forecasting whereas for the first model we would need to know, or at least estimate, next week's phone calls to forecast sales.

Notice the dw options. These request the "Durbin Watson" statistic which is a test for autocorrelation, that is, correlation between successive residuals. Autocorrelation is a commonly occurring violation of the independence assumptions when data are taken over time. The option also gives an estimate

r of the first order autocorrelation. We get dw $= 1.407$ and $\widehat{\rho} = .283$ for the first model, dw $= .969$ and $\widehat{\rho} = .497$ for the second model.

3. The Durbin-Watson statistic and first order autocorrelation.

The Durbin-Watson statistic is dw $= \sum_{t=2}^{n}(r_t - r_{t-1})^2 / \sum_{t=1}^{n} r_t^2$ where r_t is the residual at time t. If e_t represents white noise (an uncorrelated sequence) then we find these expected values:

$E\{(e_t - e_{t-1})^2\} = E\{e_t^2 - e_t e_{t-1} + e_{t-1}^2\} = \sigma^2 + 0 + \sigma^2$
$E\{e_t^2\} = \sigma^2$

Thus $\sum_{t=2}^{n}(e_t - e_{t-1})^2 / \sum_{t=1}^{n} e_t^2$ should be near 2, that is, the Durbin Watson statistic should be near 2 if calculated on a white noise sequence. If there is positive correlation between neighboring e's then e_t and e_{t-1} would be more alike than in the white noise case so that $e_t - e_{t-1}$ would be smaller in magnitude and thus dw would move toward 0.

The first order correlation in the residuals is

$$\widehat{\rho} = \sum_{t=2}^{n}(r_t - \overline{r})(r_{t-1} - \overline{r}) / \sum_{t=1}^{n}(r_t - \overline{r})^2$$

which is very close to what one would get by simply inserting r_t and r_{t-1} into the standard formula for a correlation. If, as in our example, the regression contains an intercept then $\overline{r} = 0$. It is well known that if $\widehat{\rho}$ is computed on a white noise series and if the sample size n is reasonably large, then

$$Z \; = \sqrt{n}\widehat{\rho}\sqrt{1 - \widehat{\rho}^2}$$

is approximately a $N(0, 1)$ random variable so that for large samples, values of $|Z|$ exceeding 1.96 would give us reason to suspect that autocorrelation is present.

A bit of algebra demonstrates that the Durbin-Watson statistic is roughly equal to $2(1 - \widehat{\rho})$ so that from our Z, we could get a large sample approximate distribution for the Durbin-Watson statistic. The real contribution of Durbin and Watson was to to show how to get the exact finite sample distribution of the statistic dw.

Unfortunately the Durbin-Watson theory shows that the exact distribution depends on the values of the X explanatory variables in the regression so that each new problem encountered would require a new table of critical values. However, if none of the X variables are lagged Y values and the errors are normal, they were able to calculate bounds for all critical values. Thus if you enter the tables of Durbin and Watson for a certain sample size and number of explanatory variables you will see upper and lower bounds for the true critical value.

A dw to the left of the lower bound is clearly less than the critical value and thus too close to 0 to accept the independence hypothesis (under which dw should be near 2). A dw to the right of the upper bound makes it clear that dw is closer to 2 than is the critical value so we cannot reject independence. A dw between the bounds just tells you that the calculated dw and the critical value are between these numbers so you have no idea how they are placed relative to each other.

Durbin and Watson also gave a computationally intensive way of computing p-values using the observed Xs. We will see how to get p-values from PROC AUTOREG. It should be noted that the restriction that lagged Ys not be included in the explanatory X variables still holds so that a model with lagged Ys, explicitly or implicitly among the explanatory variables, would not produce exact p-values using the Durbin-Watson method.

Because the large sample Z approximation is reasonably good, the tables of Durbin and Watson typically do not extend to very large n values so for our example residuals, we use

$Z = \sqrt{n}\widehat{\rho}/\sqrt{1-\widehat{\rho}^2}$ getting

$$\sqrt{156}(.283)/\sqrt{1-.283^2} = 3.7$$

for the first model, and

$$\sqrt{156}(.497)/\sqrt{1-.497^2} = 7.2$$

for the second model. Using 1.96 as a critical value we have strong evidence for autocorrelation in our example.

For our example we have normal residuals with homogeneous variance, but they are clearly not independent for either of our models.

4. Adjusting for autocorrelation.

Suppose we have a simple linear regression model

$$Y_t = \beta_0 + \beta_1 X_{1t} + a_t$$

where, instead of white noise e_t our error term satisfies a model such as

$$a_t = -\alpha_1 a_{t-1} - \alpha_2 a_{t-2} - \cdots - \alpha_p a_{t-p} + e_t$$

This error model is called *autoregressive of order p* where the order refers to the number of lagged a's appearing in the equation for a_t. If $p = 1$ then the first order autocorrelation coefficient from PROC REG is a reasonable estimate of α but in higher order models, the relationship between the autoregressive *coefficients* (αs) and the autocorrelations is much more convoluted.

What happens if we just ignore the autocorrelation?

(A) The estimates of the regression coefficients are still unbiased.

(B) The estimates of the regression coefficients vary more from sample to sample than do the best estimates, but still may be reasonably efficient.

(C) Estimates of standard errors for coefficients and anything computed from them (t statistics, p-values and confidence intervals for example) are biased - often badly biased.

Using our simple linear regression and an order 1 error process $a_t = -\alpha a_{t-1} + e_t$ we note that the equation holds at both times t and $t-1$ so that, multiplying through by the autoregressive parameter α we obtain

$$Y_t = \beta_0 + \beta_1 X_{1t} + a_t$$

$$\alpha Y_{t-1} = \alpha\beta_0 + \beta_1 \alpha X_{1t-1} + \alpha a_{t-1}$$

and adding we have the *transformed model*

$$(Y_t + \alpha Y_{t-1}) =$$

$$\beta_0(1+\alpha) + \beta_1(X_{1t} + \alpha X_{1t-1}) + (a_t + \alpha a_{t-1})$$

Now we note the following points about this transformed model:

(A) This is a linear model in the transformed variables (in parentheses)

(B) It has the same coefficients as the original model

(C) It has error term $(a_t + \alpha a_{t-1})$ where we are assuming $a_t + \alpha a_{t-1} = e_t$, that is this new model satisfies all the usual regression assumptions!

(D) It has $n-1$, not n observations.

Note that point (C) implies that running an ordinary regression would suffice if we knew α (or could approximate it well). We can recover the lost observation by noting that

$$\sqrt{1-\alpha^2}Y_1 = \beta_0\sqrt{1-\alpha^2} + \beta_1\sqrt{1-\alpha^2}X_{11} + a_1$$

This works because $\mathrm{Var}(a_t) = \sigma^2/(1-\alpha^2)$. What we will do is regress column 1 on columns 2 and 3 in this table using the first order autocorrelation of the residuals from an initial ordinary least squares regression as an estimate of α.

$\sqrt{1-\alpha^2}Y_1$	$\sqrt{1-\alpha^2}$	$\sqrt{1-\alpha^2}X_1$
$Y_2+\alpha Y_1$	$1+\alpha$	$X_2+\alpha X_1$
$Y_3+\alpha Y_2$	$1+\alpha$	$X_3+\alpha X_2$
$\vdots$	$\vdots$	$\vdots$
$Y_n+\alpha Y_{n-1}$	$1+\alpha$	$X_n+\alpha X_{n-1}$

These new estimates of the βs can be used to compute new residuals, a new estimate of α, new columns in the table and the whole process can be iterated until the estimates stabilize.

Alternatively, one can simply notice at the outset that this whole procedure amounts to an attempt to minimize the error sum of squares in a nonlinear model and thus use standard nonlinear search techniques (i.e. full blown maximum likelihood estimation of α and the βs simultaneously instead of the alternating technique above). Either way, we have estimated the model

$$Y_t = -\alpha Y_{t-1} + \beta_0(1+\alpha) + \beta_1(X_{1t} + \alpha X_{1t-1}) + e_t$$

which is seen to be a model that includes lagged Y's among the explanatory variables.

Now it is possible that the model needs more than 1 lagged residual. The procedure is essentially the same, we just need more terms in the transformation and more than 1 observation at the beginning needs to be recovered in a special way.

5. PROC AUTOREG for the sales data.

We use PROC AUTOREG on our sales data with the following code:

```
proc autoreg;
model y = t x/nlag=4 backstep dwprob partial;
```

The output begins with ordinary least squares, used to get the residuals and model them. Next come autocorrelations of the residuals. The lag j autocorrelation is simply an estimate of the correlation between a_t and a_{t-j} computed from the ordinary least squares residuals r_t (think of r_t as an estimate of a_t). The j^{th} partial autocorrelation can be interpreted as an estimate of the last coefficient in the regression of a_t on $a_{t-1}, \ldots, a_{t-j}$ and thus would be 0 after the appropriate number of autoregressive lags are included in the model. Time series experts use autocorrelations and partial autocorrelations to diagnose the nature of correlation in the residuals.

The drop in the autocorrelations after lag 1 is more dramatic than that in the partial autocorrelations. This suggests that a model other than autoregressive for the error terms might be considered, thus taking us into the realm of PROC ARIMA which we discuss later.

```
      Estimates of Autocorrelations

   Lag   Covariance   Correlation

    0     446.1458     1.000000
    1      126.196     0.282858
    2     -17.9134    -0.040151
    3     27.51148     0.061665
    4     8.913212     0.019978
```

Partial Autocorrelations

1	0.282858
2	-0.130610
3	0.123244
4	-0.047635

The next part of the output is a backward elimination of insignificant autoregressive parameters (α's) starting with the least significant.

Backward Elimination of Autoregressive Terms

Lag	Estimate	t-Ratio	Prob
4	0.047635	0.5821	0.5614
3	-0.123244	-1.5210	0.1304
2	0.130610	1.6188	0.1076

We are left, then, with just a lag 1 autoregressive model. The procedure next summarizes the model:

Estimates of the Autoregressive Parameters

Lag	Coefficient	Std Error	t Ratio
1	-0.28285812	0.077798	-3.636

Yule-Walker Estimates

SSE	63800.85	DFE	152
MSE	419.7425	Root MSE	20.48762
SBC	1401.124	AIC	1388.924
Reg Rsq	0.7101	Total Rsq	0.7764
Durbin-Watson	1.8850	PROB < DW	0.2013

The autoregressive coefficient -0.2829 divided by its standard error is t = -3.636 and since our sample size is reasonably large, t exceeding 1.96 in magnitude is considered significant. The error mean square 419.7 estimates σ^2, the variance of e_t so our error model is

$$a_t = .2829a_{t-1} + e_t$$

The SBC and AIC are information criteria that can be used to compare models with differing numbers of parameters. The model delivering the smallest information criterion would be selected. The criteria trade the fit of the model off against its complexity just as a person might trade the functionality of a new computer against its cost in deciding which machine to purchase.

We observe that the Durbin-Watson statistic on the transformed model is now close to 2 and an approximate p-value larger than 0.05 appears. This is not an exact p-value since the transformed model implicitly uses an estimated coefficient on lagged Y's to predict the current Y and thus does not satisfy Durbin and Watson's assumptions.

Finally there are 2 R-square values. This is because, in predicting Y one step ahead, we can use a prediction based on the X's and their coefficients only or we can add to that a forecast of the next residual based on our autoregressive error model and the most recently observed residual(s). The percentage of variation explained under these scenarios are the regression R-square and total R-square respectively. In that sense, the total R-square is a predictability R-square while the regression R-square tells how much of the predictability is associated with the X variables (which may be difficult or expensive to obtain - especially future values of them).

The procedure next uses the estimated α to get the transformed variables, e.g. $Y_t - 0.2829Y_{t-1}$, and runs ordinary least squares on the transformed variables. Because of the transformation, the resulting coefficients, standard errors, etc. are correct and, except for the fact that α is estimated instead of known, the X coefficients are the best linear unbiased estimates of the parameters.

The model parameters part of the output is given next. A portion of this is shown here:

Variable	DF	B Value	t Ratio	Approx Prob
Intercept	1	5.571991	0.882	0.3793
T	1	0.038574	0.732	0.4656
X	1	0.947411	18.220	0.0001

We see that the time trend (T) which seemed to appear in our graph does not seem important after X is included in our model. Note that this does not say that there is no increase in sales, only that there is no increase beyond what would have been predicted from incoming phone calls. Phone calls X are strongly significant. We might have been happier had the significance results been reversed, since we need to supply next week's phone volume to predict next week's sales in a model involving X.

What are our choices in terms of forecasting? One option is to somehow get estimates of future X values. Here are some forecasts of phone volume which actually came from SAS PROC ARIMA. They have been appended to our original data and you see their Y values are missing.

T	DATE	X	Y
154	12/16/94	153.000	141
155	12/23/94	146.000	159
156	12/30/94	180.000	220
157	01/06/95	157.292	.
158	01/13/95	141.705	.
159	01/20/95	131.008	.
160	01/27/95	123.665	.
161	02/03/95	118.625	.

Do we really think X will be 157.292 next week? Of course not, this is just a forecast so the use of this X value in computing a forecast for Y will add to the inaccuracy of the forecast. On the other hand, since we stopped in week 156, the use of $t = 157$ for next week's t in our model will introduce no inaccuracy in the forecast.

Choosing SBC as a criterion we have

Model	SBC	MSE	α
X and t	1401	420	-.2829
X only	1397	418	-.2873
t only	1607	1658	-.4972

so the model with X only is clearly preferred.

One confusing phenomenon that often occurs is that, although statistical theory indicates that the estimates from our transformed model should be better than the ordinary least squares estimates that ignore autocorrelation, the ordinary least squares printout shows smaller standard errors than the ones shown in PROC AUTOREG. How can that be if the PROC AUTOREG method provides better estimates? The answer is simple - the ordinary least squares numbers are not good estimates of the standard errors and are often, but not always, deceptively small. In other words by ignoring autocorrelation we are using inferior estimates of the parameters but the standard errors falsely indicate that they are superior.

The three models estimated by PROC AUTOREG (with standard errors) are:

$$\begin{aligned} Y_t = {} & \underset{(6.3191)}{5.5720} + \underset{(0.0527)}{0.0386t} + \underset{(0.052)}{0.9474X_t} \\ Y_t = {} & \underset{(5.8111)}{7.3681} + \underset{(0.0498)}{0.9589X_t} \\ Y_t = {} & \underset{(11.0956)}{83.8472} + \underset{(0.1222)}{0.3388t} \end{aligned}$$

6. Forecasting

The model with only X is

$$Y_t = \underset{(5.8111)}{7.3681} + \underset{(0.050)}{.9589X_t} + a_t$$

with $a_t = .2873 \quad a_{t-1} + e_t$ and we are at the last week of year 3 in our data so t=156. Standard errors are in parentheses. The last observation was $Y_{156} = 220$ and $X_{156} = 180$ so that the residual, an estimate of a_{156}, would be $r_t = 220 - 7.3681 - .9589(180) = 40.030$ and we predict a_{157} as $.2873(40.030) = 11.501$. Now if we assume X_{157} will be 157.292 then we predict Y_{157} to be $7.3681 + .9589(157.292) + 11.501 = 158.195 + 11.501 = 169.696$. This is the first forecast in the PROC AUTOREG output dataset and an associated standard error 20.9 is used to compute a 95% prediction interval. The problem is that this standard error is computed assuming that X will be exactly 157.292 in the next time interval. Our true level of uncertainty in the forecast of Y would certainly be influenced by the variability in our imputed X value.

7. Using PROC ARIMA

We can model the whole process, X and Y, in PROC ARIMA. We first compute a model for X which PROC ARIMA estimates as

$$X_t - 107.6 = .6864(X_{t-1} - 107.6) + e_t$$

Now PROC ARIMA can fit and forecast the same model as PROC AUTOREG, however it gives you the option of using the X model to provide forecasts of future X's and it incorporates the uncertainty in those X's in the Y forecasts. Our forecasts of future X's were from PROC ARIMA so the forecasts from the two procedures will match but the forecast error variances will differ. We close by presenting graphs of forecasts and their standard errors from several scenarios.

In figure 5, the forecasts from PROC AUTOREG and PROC ARIMA with their intervals are overlaid. The difference in widths of the forecast intervals illustrates the magnitude of error that is being ignored when one treats future X's as known when they are in fact forecasts.

Figure 6 shows the forecasts and intervals from the model that uses time t as the explanatory variable. Here we can properly treat future t's as known, but pay a price in that the model does not fit as well. The forecasted X ARIMA plot is overlaid on this (it gives the lower forecasts and intervals) and it is interesting to note that although this model fit substantially worse according to our statistical tests, once we admit that there is error in our forecasts of X, our forecasts and error bands are similar in both models. In the long term, the model with t will give forecasts that trend linearly upward while the forecasts with the ARIMA model will return to the historic mean as will always happen with a stationary ARIMA model.

Finally, PROC ARIMA allows the fitting of a larger class of error models than autoregressive. A lag 1 moving average model also provides an excellent fit to the error series for the sales data. Using the autoregressive model for X, a linear relationship between Y and X, and an order 1 moving average error term, our estimated model becomes

$$Y_t = 8.607 + 0.95\ X_t + e_t + .37e_{t-1}$$

$$X_t - 107.6 = 0.686\ (X_{t-1} - 107.6) + u_t$$

Because the error term $e_t - \theta e_{t-1}$ is a moving average we estimated this model in PROC ARIMA:

```
proc arima data=a;
i var=x noprint; e p=1 ml;
i var=y crosscor=(x) nlag=5;
e input = (x) q=1 ml;
f lead=5 out=out5 id=t;
```

Because our input variable is modeled, we will get a crosscorrelation plot which has been prewhitened. It is a plot of correlations between Y at time t and X at time $t - j$ which has been cleared of any indirect correlations caused by autocorrelation in the X series. It is thus a picture of how changes in X are incorporated into the Y series.

The plot of the crosscorrelatins for our example will look similar to this:

```
              Crosscorrelations
Lag   Corr
 -5    -0.065        .   *    |      .
 -4     0.043        .        |  *   .
 -3    -0.090        .  **    |      .
 -2    -0.060        .    *   |      .
 -1     0.008        .        |      .
  0     0.797        .        |  **********
  1     0.057        .        |  *   .
  2    -0.010        .        |      .
  3    -0.053        .   *    |      .
  4     0.135        .        |  ***
  5    -0.124        .  **    |      .
```

“.” marks two standard errors

It is seen that there is no lagged correlation, only contemporaneous correlation. The moving average error structure gave error mean square 408.7 as compared to 418.3 so it is the best fit of all models considered here by that criterion.

SAS is the registered trademark of SAS Institute Inc. in the USA and other countries ™ indicates USA registration.