Fitting time series regression models

Why do simple time series models sometimes outperform regression models fitted to nonstationary data?

- Two nonstationary time series X and Y generally don't stay perfectly "in synch" over long periods of time--i.e., they do not usually maintain a perfectly linear relationship--even if they are causally related.

- There may be some "omitted variable", say Z, which could in principle explain some of the discrepancy in the relationship between X and Y--but this is not the only possibility. For example, the nature of the relationship between X and Y may simply change over time.

Remember that if X and Y are *nonstationary*, this means that we cannot necessarily assume their statistical properties (such as their correlations with each other) are constant over time.

- If the variables X and Z change relatively *slowly* from period to period, it is possible that the information X(t) *and* Z(t) contain with respect to Y(t) is already contained in Y(t-1), Y(t-2), etc.--i.e., in Y's own recent history--which time series models are able to utilize.

In other words, recent values of Y might be good "proxies" not only for the effect of X *but also for the effects of any omitted variables*.

- Business and macroeconomic times series often have strong *contemporaneous* correlations, but significant *leading* correlations--i.e., cross-correlations with other variables at positive lags--are often hard to find. Thus, regression models may be better at predicting the present than the future.

How to get the best of both worlds--regression and time series models:

1. *Stationarize* the variables (by differencing, logging, deflating, or whatever) before fitting a regression model.

- If you can find transformations that render the variables stationary, then you have greater assurance that the correlations between them will be stable over time.

- Stationarizing also implicitly brings the recent history of the variables into the forecast.

Example: instead of regressing Y on X, regress DIFF(Y) on DIFF(X)

The regression equation is now: **$\tilde{Y}(t) - Y(t-1) = a + b(X(t) - X(t-1))$**

...which is equivalent to : **$\tilde{Y}(t) = Y(t-1) + a + bX(t) - bX(t-1)$**

Notice that this brings the prior values of *both* X and Y into the prediction.

---

2. Use *lagged* versions of the variables in the regression model.

- This allows varying amounts of recent history to be brought into the forecast
- Lagging of independent variables is often necessary in order for the regression model to be able to predict the future--i.e., to predict what will happen in period t based on knowledge of what happened up to period t-1

Example: instead of regressing Y on X, regress Y on LAG(X,1) and LAG(Y,1)

The regression equation is now **$\tilde{Y}(t) = a + bX(t-1) + cY(t-1)$**

---

3. It often helps to do *both*--i.e., to stationarize the variables by differencing, then use lags of the stationarized variables.

Example: regress DIFF(Y) on LAG(DIFF(X),1) and LAG(DIFF(Y),1)

The regression equation is now: **$\tilde{Y}(t) - Y(t-1) = a + b(X(t-1) - X(t-2)) + c(Y(t-1) - Y(t-2))$**

...which is equivalent to **$\tilde{Y}(t) = Y(t-1) + a + b(X(t-1) - X(t-2)) + c(Y(t-1) - Y(t-2))$**

How to decide which combinations of lags or differences to try?

1. Determine which transformations (if any) are needed to stationarize each variable by looking at **time series plots** and **autocorrelation plots**

- Time series plots of stationary variables should have a *well-defined mean* and a relatively *constant variance* (i.e., no heteroscedasticity).
- The autocorrelations of a *nonstationary* variable will be *strongly positive* and *non-noisy*-looking out to a high number of lags (often 10 or more)--i.e., they will be "slow to decay"
- The autocorrelation plot of a *stationary* variable will usually decay into "noise" and/or hit negative values within 3 or 4 lags.

- If the *lag-1 autocorrelation* is already *negative*, no more differencing is needed. If the lag-1 autocorrelation is -0.5 or more negative, the variable may have been "overdifferenced."

---

2. Look at **autocorrelations** of the stationarized dependent variable (e.g., DIFF(Y)) to determine whether one or more of its lags is likely to be helpful

- Concentrate mainly on what happens at the *first few lags* and (in the case of seasonal data) what happens at the *seasonal lag* (e.g., at lag 12 for monthly data).

- "Spikes" in the autocorrelation plot at *peculiar* lags (e.g. lag 5 or 7 in a monthly time series) are probably mere "noise" caused by a chance alignment of extreme values.

---

3. Look at **cross-correlations** between the stationarized dependent variable (the "first" series) and stationarized independent variables (the "second" series).

- A significant cross-correlation at a **positive lag** indicates that the independent variable may be significant when lagged by that number of periods.

- For example, if DIFF(X) is the second time series and a significant cross-correlation is observed at lag 1, this suggests that LAG(DIFF(X),1) might be a significant predictor of the dependent variable.

- The cross-correlation function, like the autocorrelation function, is typically noisy. Cross-correlations at lags of 3 or more are often merely accidental (except where seasonal effects are important). Generally you should take seriously only the cross-correlations at lags 0, 1, and 2.

- The cross-correlation function, like the autocorrelation function, will typically show a smooth pattern of significantly positive correlations out to a high number of lags if the variables have *not* been properly stationarized--but this doesn't mean much.

Fitting a regression model to differenced and/or lagged data:

1. Regress the stationarized dependent variable on lags of itself and/or stationarized independent variables as suggested by autocorrelation and cross-correlation analysis

Example: DIFF(Y) shows a significant autocorrelation at lags 1 and 2 but not at higher lags, and DIFF(Y) shows a significant cross-correlation with DIFF(X) at lags 0 and 1.

- Try regressing DIFF(Y) on LAG(DIFF(Y),1), LAG(DIFF(Y),2), and LAG(DIFF(X),1).

- You could also include DIFF(X) as indicated by the cross-correlation at lag 0, but this would preclude being able to predict one period ahead.

- If your dependent variable is DIFF(Y), then the forecast reports and graphs produced by your model will all be for *differences* of Y, not for Y in its original units. The regression procedure does not automatically "undifference" the output for you.

(If you wish to print or plot undifferenced forecasts, you will need to use spreadsheet commands either in Statgraphics or Excel to add the predicted differences to the previous actual values in order to get predictions for Y itself in each period.)

- For stock price data, on which you would probably use a DIFF(LOG( )) transformation, obtaining results in differenced terms is not necessarily bad: it is more interesting to plot the predicted *returns* rather than predicted prices, so you can see if the predicted returns deviate significantly from a constant.

---

2. You can use a "stepwise" approach to model-fitting but beware of over-fitting the data. Be cautious in choosing how many lags and how many different independent variables to include at the beginning of the process.

- The Multiple Regression procedure includes automatic stepwise regression as a right-mouse-button option. (The "forward" method is usually safer.)

- The Advanced Regression module contains an all-possible-regressions option ("Regression Model Selection" procedure)-- but beware!

---

3. It is especially important to VALIDATE your model with hold-out data when selecting models by automatic methods. Ideally you should withhold data during the model-selection process as well as during the final testing of the model.

- The Multiple Regression procedure does not offer any options for validation, alas.

- The Advanced Regression module includes a validation option in all its procedures: use the "Select" field to hold out data--e.g., INDEX<81 to hold out all data after row 80.

---

4. The good news: The Forecasting procedure can be used to fit regression models to lagged and differenced data *and* to validate them.

- Differencing and lagging of the dependent variable can be performed by using ARIMA options.

- For example, to use DIFF(Y) as the dependent variable and include LAG(DIFF(Y),1)) and LAG(DIFF(Y),2) as regressors, you would just use Y as the input variable, then specify an ARIMA model with 1 nonseasonal difference and set AR=2. Forecasts will be automatically "undifferenced" in the output reports and graphs.

(Setting AR=2 means that you want 2 *autoregressive* terms--i.e., the first two lags of the differenced dependent variable--included in the forecasting equation.)

---

5. The bad news: You can also, in principle, add *regressors* to the ARIMA model such as LAG(DIFF(X),1) etc. However, the regression option gives a "data error" message if there is more than 1 missing value at the beginning of a regressor, which will be the case if the total number of lags or differences is more than 1. (The regressor could either be lagged by one period, or else differenced, but not both.)

- **Workaround:** use the "Generate Data" command to create new columns on your data spreadsheet containing all lagged and/or differenced variables that may be used as regressors, and assign them short descriptive names. Then *delete any rows at the beginning of the file that contain missing values of these variables*. In the Forecasting procedure, specify the regressor variables *by name* rather than using expressions that involve LAG or DIFF.