How to choose forecasting models

Steps in choosing a forecasting model
Forecasting flow chart
Data transformations and forecasting models: what to use and when
Automatic forecasting software
Political and ethical issues in forecasting
How to avoid trouble: principles of good data analysis

Data transformations and forecasting models: what to use and when

---

Transformation	Properties	When to use	Points to keep in mind
Deflation by CPI or another price index	Converts data from nominal dollars (or other currency) to constant dollars; usually helps to stablilize variance	When data are measured in nominal dollars (or other currency) and you want to explicitly show the effect of inflation--i.e., uncover "real growth"	To generate a true forecast for the future in nominal terms, you will need to make an explicit forecast of the future value of the price index--i.e., you will need to forecast the inflation rate (but this is easy if you're in a period of steady inflation)
Deflation at a fixed rate	Merely applies a constant discount factor to past data	When you only need to approximately model the effect of past inflation and/or you wish to impose an assumption about the current and future inflation rate--you can twiddle the inflation rate to see what value does the best job of flattening out the trend and/or stabilizing the variance	When used with a zero-trend model like simple exponential smoothing or random walk without growth, the assumed inflation rate is precisely the percentage growth in the future forecasts.
Logarithm	Converts *multiplicative* patterns to *additive* patterns and/or linearizes exponential growth; converts *absolute* changes to *percentage* changes; often stablizes the variance of data with compound growth, regardless of whether deflation is also used	When compound growth is not due to inflation (e.g. when data is not measured in currency); when you do not need to separate inflation from real growth; when data distribution is positive and highly skewed (e.g., exponential or log-normal distribution); when variables are multiplicatively related	Logging is *not the same* as deflating: it linearizes growth but does not remove a general upward trend; if logged data still have a consistent upward trend, then you should use a model that includes a trend factor (e.g., random walk with growth, ARIMA, linear exponential smoothing).
First difference	Converts "levels" to "changes"	When you need to stationarize a series with a strong trend and/or random-walk behavior (often useful when fitting regression models to time series data)	Differencing is an explicit option in ARIMA modeling and it is implicitly a part of random walk and exponential smoothing models; therefore you would *not* manually difference the input variable (using the DIFF function) when specifying model type as "random walk" or "exponential smoothing" or "ARIMA"; first difference of LOG(Y) is the *percentage* change in Y
Seasonal difference	Converts "levels" to "seasonal changes"	When you need to remove the gross features of seasonality from a strongly seasonal series without going to the trouble of estimating seasonal indices	Seasonal differencing is an explicit option in ARIMA modeling; you MUST include a seasonal difference (as a modeling option, *not* an SDIFF transformation of the input variable) if the seasonal pattern is consistent and you wish it to be maintained in long-term forecasts
Seasonal adjustment	Removes a constant seasonal pattern from a series (either multiplicative or additive)	When you wish to separate out the seasonal component of a series and then fit what's left with a nonseasonal model (regression, smoothing, use the multiplicative version unless data has been logged	Adds a lot of parameters to the model--one for each season of the year. (In Statgraphics, the seasonal indices are not explicitly shown in the output of the Forecasting procedure--you must separately run the Descriptive Methods procedure to display the seasonal indices.)

Model type	Properties	When to use	Points to keep in mind
Random walk	Predicts that "next period equals this period" (perhaps plus a constant); a.k.a. ARIMA(0,1,0) model	As a baseline against which to compare more elaborate models; when applied to logged data, it is a "geometric" random walk--the default model for stock market data	Plot of forecasts looks exactly like a plot of the data, except lagged by one period (and shifted slightly up or down if a growth term is included); long term forecasts follow a straight line (horizontal if no growth term is included); confidence intervals for long-term forecasts widen according to a square-root law (sideways-parabola shape); logically equivalent to MEAN model fitted to DIFF(Y)
Linear trend	Regression of Y on the time index	Rarely the best model for forecasting--use only when you have *very* few data points and no obvious pattern in data other than a slight trend; can be used in conjunction with seasonal adjustment--but if you have enough data to seasonally adjust, you probably should use another model	Forecasts follow a straight line whose slope equals the average slope over the whole estimation period but whose intercept is anchored in the distant past; short-term forecasts therefore may miss badly and confidence intervals for long-term forecasts are usually not reliable; other models that extrapolate a linear trend into the future (random walk with growth, linear exponential smoothing, ARIMA models with 1 difference w/constant or 2 differences w/o constant) often do a better job by "reanchoring" the trend line on recent data
Simple moving average	Simple (equally weighted) average of recent data	When data are in short supply and/or highly irregular	Primitive but relatively robust against outliers and messy data; long-term forecasts are a horizontal line extrapolated from the most recent average; a long-term trend can be incorporated via fixed-rate deflation at an assumed interest rate
Simple exponential smoothing	Exponentially weighted average of recent data; "average age" of data in forecast (amount by which forecasts lag behind turning points) is 1/alpha; same as an ARIMA(0,1,1) model without constant	When data are nonseasonal (or deseasonalized) and display a time-varying mean without a consistent trend	Long-term forecasts are a horizontal line extrapolated from the most recent smoothed value; same as a random walk model without growth if alpha=0.9999; forecasts get smoother and slower to respond to turning points as alpha approaches zero; confidence intervals widen less rapidly than in the random walk model; a long-term trend can be incorporated via fixed-rate deflation at an assumed interest rate or by fitting an ARIMA(0,1,1) model *with* constant
Linear exponential smoothing (Brown's or Holt's)	Assumes a time-varying linear trend as well as a time-varying level (Brown's uses 1 parameter, Holt's uses separate smoothing parameters for level and trend); essentially an ARIMA(0,2,2) model without constant	When data are nonseasonal (or deseasonalized) and display time-varying local trends (usually applicable to data that are "smoother" in appearance--i.e., less noisy--than what would be well fitted by simple exponential smoothing)	Long-term forecasts follow a straight line whose slope is the estimated local trend at the end of the series; confidence intervals for long-term forecasts widen rapidly--the model assumes that the future is VERY uncertain because of time-varying trends; often does not outperform simple exponential smoothing, even for data with trends, because extrapolation of time-varying trends is risky
Seasonal random walk	Predicts that "next period equals same period last year" (plus constant); an ARIMA(0,0,0)x(0,1,0) model with constant	As a baseline against which to compare fancier seasonal models; as foundation for seasonal ARIMA models (e.g., (1,0,0)x(0,1,1))	Long-term forecasts have same seasonal pattern as last year; long-term trend is equal to the *average* trend over whole past history of series; confidence intervals widen slowly; slow to respond to cyclical upturns and downturns; logically equivalent to MEAN model fitted to SDIFF(Y,s)
Seasonal random trend	Predicts that change from this period to next period will be the same as change observed at this time last year; an ARIMA(0,1,0) x(0,1,0) model without constant	As a baseline against which to compare fancier seasonal models; as foundation for seasonal ARIMA models (e.g., (0,1,1)x(0,1,1) without constant)	Long-term forecasts have same seasonal pattern as last year; long-term trend is equal to the *most recently observed* annual trend; confidence intervals widen rapidly; quick to respond to cyclical upturns and downturns; logically equivalent to MEAN model fitted to DIFF(SDIFF(Y)) (with no constant--i.e., mean is assumed to be zero)
Winter's seasonal smoothing	Assumes time-varying level, trend, and seasonal indices (either	When data are trended and seasonal and you wish to decompose it into *local* level/trend/seasonal	Initialization of seasonal indices and joint estimation of three smoothing parameters is sometimes tricky--watch to see that parameter estimates converge and that

	multiplicative or additive seasonality)	factors; normally you use the multiplicative version unless data is logged	forecasts and confidence intervals look reasonable; a popular choice for "automatic" forecasting because it does a little of everything, but has a lot of parameters and sometimes overfits the data or is unstable
Multiple regression	A general linear forecasting equation involving other variables	When data are correlated with other explanatory or causal variables (e.g., price, advertising, promotions, interest rates, indicators of general economic activity, etc.); the key is to choose the right variables and the right *transformations* of those variables to justify the assumption of a linear model and to take into account the time dimension in the data	Forecasts cannot be extrapolated into the future unless and until values are available for the independent variables; for this reason the independent variables must often be lagged by one or more periods--but when *only* lagged variables are used, a regression model may fail to outperform a time series model which relies only on the history of the original series; regressions of *nonstationary* variables often have high "R-squared" but poor performance compared to time series models; it often helps to stationarize the dependent variable and/or add lags of the dependent and independent variables to the model; "automatic" model selection techniques such as stepwise regression and all-possible regressions are available, but beware of overfitting; it is important to validate the model by testing it on hold-out data and by computing its "effective R-squared" (percent of variance explained) relative to a random walk model or other appropriate time series model
ARIMA	A general class of models that includes random walk, random trend, seasonal and non-seasonal exponential smoothing, and auto-regressive models; forecasts for the stationarized dependent variable are a linear function of lags of the dependent variable and/or lags of the errors	When data are relatively plentiful (4 seasons or more) and can be satisfactorily stationarized by differencing and other mathematical transformations; when it is not necessary to explicitly separate out the seasonal component (if any) in the form of seasonal indices	ARIMA models are designed to squeeze all autocorrelation out of the original time series; a systematic procedure exists for identifying the best ARIMA model for any given time series; features of ARIMA and multiple regression models can be combined in a natural way; ARIMA models often provide a good fit to highly aggregated, highly plentiful data; they may perform relatively less well on disaggregated, irregular, and/or sparse data