# Forecasting with moving averages

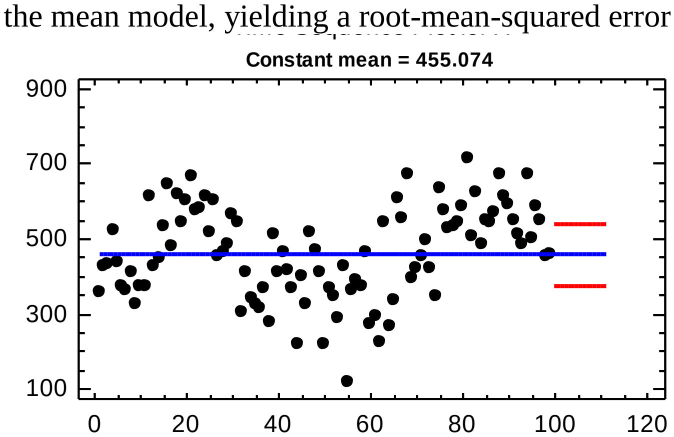
Robert Nau
Fuqua School of Business, Duke University
August 2014

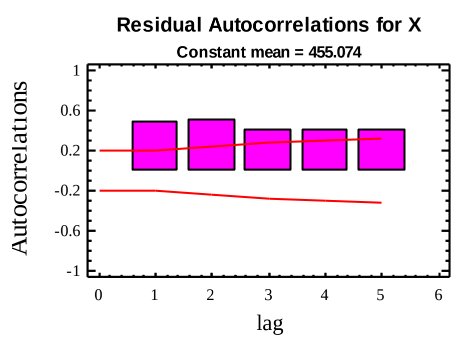
1. Simple moving averages
2. Comparing measures of forecast error between models
3. Simple exponential smoothing
4. Linear exponential smoothing
5. A real example: housing starts revisited
6. Out-of-sample validation

## 1. SIMPLE MOVING AVERAGES

In previous classes we studied two of the simplest models for predicting a model from its own history—the mean model and the random walk model. These models represent two extremes as far as time series forecasting is concerned. The mean model assumes that the best predictor of what will happen tomorrow is the average of everything that has happened up until now. The random walk model assumes that the best predictor of what will happen tomorrow is what happened today, and all previous history can be ignored. Intuitively there is a spectrum of possibilities in between these two extremes. Why not take an average of what has happened in some window of the recent past? That's the concept of a "moving" average.

You will often encounter time series that appear to be "locally stationary" in the sense that they exhibit random variations around a local mean value that changes gradually over time in a non-systematic way. Here's an example of such a series and the forecasts that are produced for it by the mean model, yielding a root-mean-squared error (RMSE)[1] of 121:

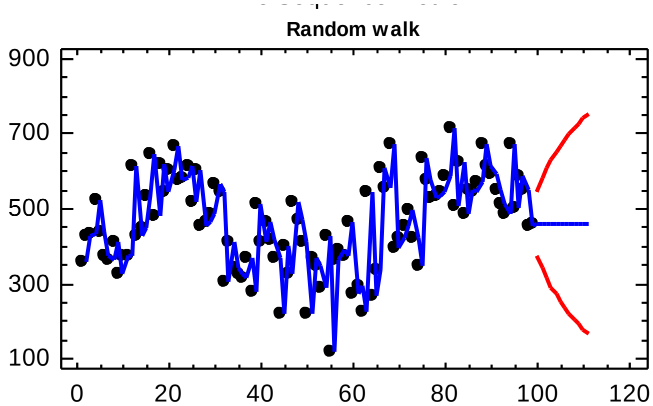


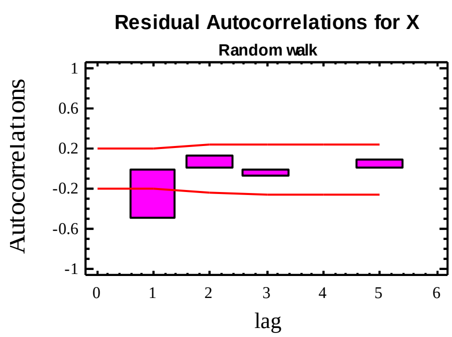




1 The mean squared error (MSE) statistic that is reported in the output of various statistical procedures is *the simple average of the squared errors*, which is equal to the population variance of the errors plus the square of the mean error, and RMSE is its square root. RMSE is a good statistic to use for comparing models in which the mean error is not necessarily zero, because it penalizes bias (non-zero mean error) as well as variance. RMSE does not include any adjustment for the number of parameters in the model, but very simple time series models usually have at most one or two parameters, so this doesn't make much difference.

Here the local mean value displays a cyclical pattern. The (global) mean model doesn't pick this up, so it tends to overforecast for many consecutive periods and then underforecast for many consecutive periods. This tendency is revealed in statistical terms by the autocorrelation plot of the residuals (errors). We see a pattern of strong positive autocorrelation that gradually fades away, rather than a random pattern of insignificant values. In particular, the autocorrelations at lags 1 and 2 are both around 0.5, which is far outside the 95% limits for testing a significant departure from zero (the red bands). The 50% (not 95%) confidence limits for the forecasts are also shown on the time series plot, and they are clearly not realistic. If the model is obviously wrong in its assumptions, then neither its point forecasts nor its confidence limits can be taken seriously.

Now let's try fitting a random walk model instead. Here are the forecasts 50% limits, and residual autocorrelations:




At first glance this looks like a much better fit, but its RMSE is 122, about the same as the mean model. (122 is not "worse" than 121 in any practical sense. You shouldn't split hairs that finely.) If you look closer you will see that this model perfectly tracks each jump up or down, but it is always *one period late* in doing so. This is characteristic of the random walk model, and sometimes it is the best you can do (as in the case of asset prices), but here it seems to be *over-responding t*o period-to-period changes and doing more zigging and zagging than it should. In the residual autocorrelation plot we see a highly significant "negative spike" at lag 1, indicating that the model tends to make a positive error following a negative error, and vice versa. This means the errors are not statistically independent, so there is more signal that could be extracted from the data. The 50% confidence limits for the forecasts are also shown, and as is typical of a random walk model they widen rapidly for forecasts more than 1 period ahead according to the square-root-of-time rule.[2] Here they are *too* wide—the series appears to have some "inertia" and does not change direction very quickly. Again, if the model assumptions appear to be wrong, its confidence limits don't reflect the true uncertainty about the future.

It's intuitively plausible that a *moving*-average model might be superior to the mean model in adapting to the cyclical pattern and also superior to the random walk model in not being too sensitive to random shocks from one period to the next. There are a number of different ways in

[2] 95% limits would be three times as wide and way off the chart!

which a moving average might be computed, but the most obvious is to take a simple average of the most recent m values, for some integer m. This is the so-called *simple moving average model* (SMA), and its equation for predicting the value of Y at time t+1 based on data up to time t is:

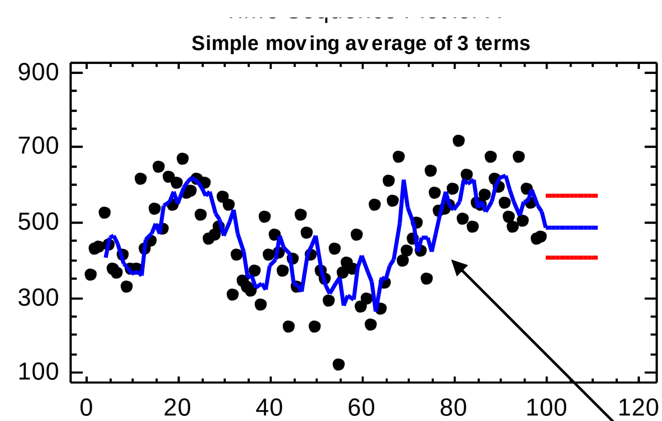
$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m}$$

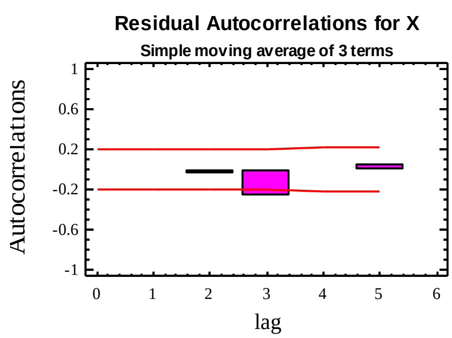
The RW model is the special case in which m=1. The SMA model has the following characteristic properties:

- ☞ Each of the past m observations gets a weight of 1/m in the averaging formula, so as m gets larger, each individual observation in the recent past receives less weight. This implies that larger values of m will filter out more of the period-to-period noise and yield *smoother-looking* series of forecasts.

- ☞ The first term in the average is “1 period old” relative to the point in time for which the forecast is being calculated, the 2nd term is two periods old, and so on up to the mth term. Hence, the “average age” of the data in the forecast is (m+1)/2. *This is the amount by which the forecasts will tend to lag behind in trying to follow trends or respond to turning points*. For example, with m=5, the average age is 3, so that is the number of periods by which forecasts will tend to lag behind what is happening now.

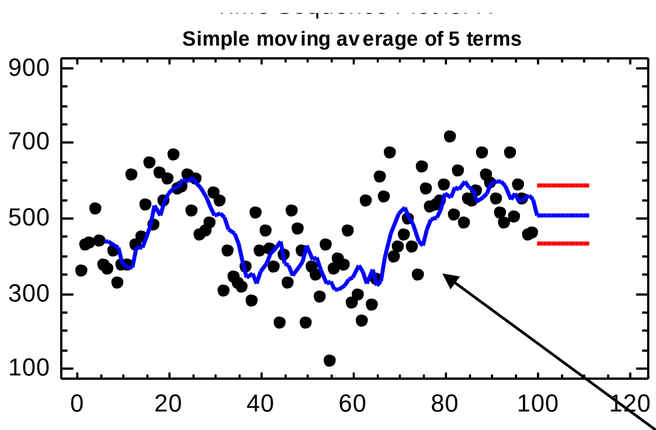
In choosing the value of m, you are making a tradeoff between these two effects: filtering out more noise vs. being too slow to respond to trends and turning points. The following sequence of plots shows the forecasts, 50% limits, and residual autocorrelations of the SMA model for m = 3, 5, 9, and 19. The corresponding average age factors are 2, 3, 5, and 10. If you look very closely, you’ll see that the forecasts of the models tend to lag behind the turning points in the data by exactly these amounts. Notice as well that the forecasts get much smoother-looking and the errors become more positively autocorrelated for higher values of m.[3]

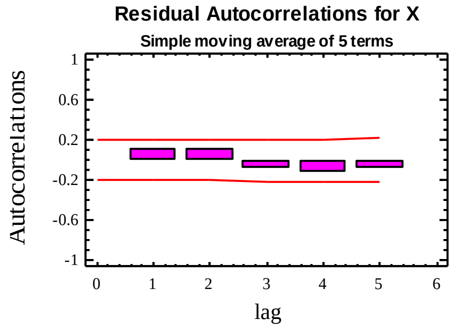


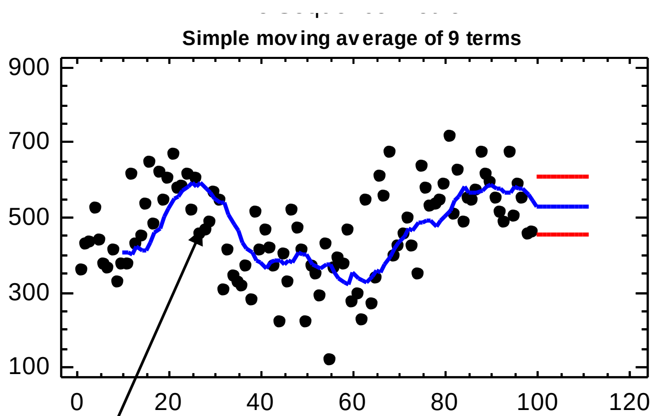

With m=3 the plot of SMA forecasts is quite choppy

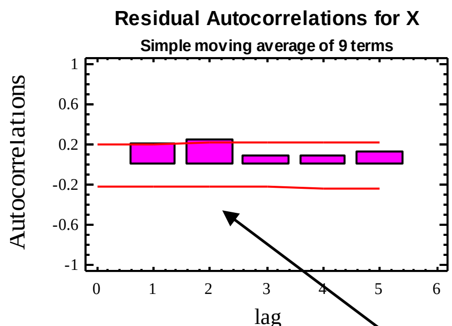
[3] The oddball negative spike at lag 3 in the 3-term model is of no consequence unless we have some a priori reason to believe there is something special about a 3-period time lag. What we are concerned with here is whether there is significant autocorrelation at the first couple of lags and whether there is some kind of overall pattern in the autocorrelations. In any case, *residual autocorrelations are not the bottom line*, just a red flag that may wave to indicate that there may be a better model out there somewhere.



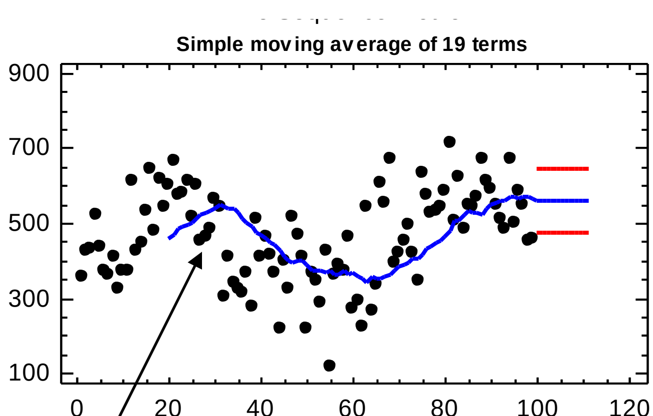



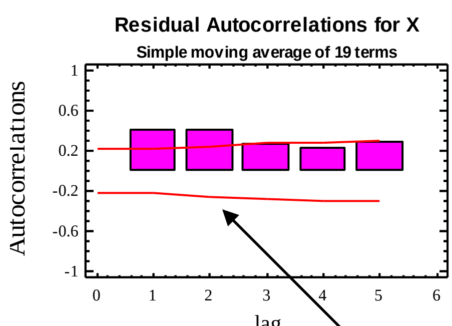

With m=5 it looks a little smoother




With m=9 the forecasts are even smoother but starting to lag behind turning points noticeably—the average age of data in the forecast is 5. The errors are also starting to be positively autocorrelated.




With m=19 the forecasts have a nice smooth cyclical pattern but they lag behind turning points by 10 periods, alas.

There is now very significant positive autocorrelation in the errors, indicating long runs of consecutive errors with the same sign.

## 2. COMPARING MEASURES OF FORECAST ERROR BETWEEN MODELS

What's the best value of m in the simple moving average model? A *good* value is one that yields small errors and which otherwise makes good sense in the decision-making environment in which it will be used. In the Forecasting procedure in Statgraphics there is a nifty (if I do say so myself) *model-comparison report* that lets you make side-by-side comparisons of error stats for *1-step-ahead forecasts* for up to 5 different time series models, which could be SMA models with different values of m or different types of models altogether. Here is the model comparison table for the random walk model and the four SMA models shown above:

**Model Comparison**
Data variable: X
Number of observations = 99
Start index = 1.0
Sampling interval = 1.0

**Models**
(A) Random walk
(B) Simple moving average of 3 terms
(C) Simple moving average of 5 terms
(D) Simple moving average of 9 terms
(E) Simple moving average of 19 terms

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 121.759 | 93.2708 | 23.6152 | 1.04531 | -5.21856 |
| (B) | **104.18** | **80.5662** | **20.2363** | 1.12125 | -5.20793 |
| (C) | **101.636** | **80.6686** | **20.2747** | 1.35328 | -5.32013 |
| (D) | **104.049** | **80.2773** | **20.1534** | 6.89349 | -4.66414 |
| (E) | 118.021 | 92.5751 | 23.6812 | 6.91815 | -6.43064 |

Models B, C, and D (m=3, 5, 9) have about equally good error stats

The various error stats are as follows:

**RMSE: root mean squared error**[4] (the most common standard of goodness-of-fit, penalizes big errors relatively more than small errors because it squares them first; it is approximately the standard deviation of the errors if the mean error is close to zero)

**MAE: mean absolute error** (the average of the absolute values of the errors, more tolerant of the occasional big error because errors are *not* squared)

**MAPE: mean absolute percentage error** (perhaps better to focus on if the data varies over a wide range due to compound growth or inflation or seasonality, in which case you may be more concerned about measuring errors in percentage terms)

**ME: mean error** (this indicates whether forecasts are biased high or low—should be close to 0)

**MPE: mean percentage error** (ditto in percentage terms)

---

[4] For a regression model, the RMSE is *almost* the same thing as the standard error of the regression—the only difference is the minor adjustment for the number of coefficients estimated. In calculating the RMSE of a forecasting model, the sum of squared errors is divided by the sample size, n, before the square root is taken. In calculating the standard error of the regression, the sum of squared errors is divided by n-p, where p is the number of coefficients estimated, including the constant. If n is large and p is small, the difference is negligible. So, focusing on RMSE as the bottom line is the same thing as focusing on the standard error of the regression as the bottom line.

Usually the best measure of the *size* of a typical error is the RMSE, provided that the errors are approximately normally distributed and that you worry about making a few big mistakes more than you worry about making a lot of little ones. At any rate, you software assumes that this is what you want to minimize, because it estimates the model parameters by the least-squares method.[5] However, MAE and MAPE are easier for non-specialists to understand, so they might be useful numbers for a presentation. They are also less sensitive to the effects of big outliers and so might give a better estimate of the size of an "average" error when the distribution of errors is far from normal. Also, MAPE gives relatively more weight to accuracy in predicting *small values* because it is computed in percentage terms. ME and MPE are usually not very important, because bias (a non-zero value for the average error) is usually small when parameters are estimated by minimizing squared error.[6] However, if there is a consistent upward or downward trend in the data, then models that do not include a trend component (like the SMA model) will have biased forecasts. Here the mean error is very slightly positive, because there is a very slight positive trend.

Models B, C, and D are very similar on all the error-size statistics, and their residual autocorrelation plots are OK (although a bit of positive autocorrelation is creeping in when you hit m=9). Only models A and E are obviously bad in terms of error measures and residual autocorrelations. C is a little better than B or D in terms of RMSE, but the difference is hairsplitting. *You don't need to go with the model whose RMSE is the absolute lowest in a case like this—you can exercise some discretion based on other considerations*. In this case there are qualitative differences between these 3 models that perhaps also should be considered. Model B (m=3) is the most responsive to the last few data points (which might be a good thing) but it also does the most zigging and zagging (which might be a bad thing). Model D (m=9) is much more conservative in revising its forecasts from period to period, and model C (m=5) strikes a balance between the two.

The SMA model can be easily customized in several ways to fine-tune its performance. If there is a consistent *trend* in the data, then the forecasts of any of the SMA models will be *biased* because they do not contain any trend component. The presence of a trend will tend to give an edge to models with lower values of m, regardless of the amount of noise that needs to be filtered out. You can fix this problem by simply *adding a constant to the SMA forecasting equation*, analogous to the drift term in the random-walk-with-drift model:

$$\hat{Y}_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m} + d$$

Simple Moving Average with Trend

[5] The simple moving average model does not have any continuous-ranged parameters for Statgraphics to estimate for you, but you can do your own estimation on the basis of RMSE (or other criteria) by manual means, as done here.
[6] ME is likely to be significantly different from zero only in cases where a log or deflation transformation has been used, with model parameters estimated by minimizing the squared error in *transformed* units. In Statgraphics, if you specify a log transformation or fixed-rate deflation transform as a model option inside the Forecasting procedure, this is what is done, and you shouldn't be surprised if you end up with some bias in the untransformed forecasts. If you use a log transformation, then you are implicitly minimizing mean squared *percentage* error, so you should expect MAPE to be relatively lower and MPE to be relatively closer to zero than without a log transformation.

Another way to fine-tune the SMA model is to use a *tapered moving average* rather than an equally-weighted moving average. For example, in the 5-term moving average model, you could choose to put only half as much weight on the newest and oldest values, like this:

$$\hat{Y}_{t+1} = \frac{\frac{1}{2}Y_t + Y_{t-1} + Y_{t-2} + Y_{t-3} + \frac{1}{2}Y_{t-4}}{4}$$

Tapered Moving Average (5-term)

This average is centered 3 periods in the past, like the 5-term SMA model, but when an unusually large or small value is observed, it doesn't have as big an impact when it first arrives or when it is finally dropped out of the calculation, because its weight is ramped up or down over two periods. So, a tapered moving average is more robust to outliers in the data.

## 3. SIMPLE EXPONENTIAL SMOOTHING

The SMA model is an easy-to-understand method for estimating the local mean value around which a time series is thought to be randomly varying, but putting equal weight on the last m observations and no weight on any previous observations is usually not the best way to average values that are arriving *consecutively in time*. Intuitively, all the past values have some relevance, but each newer one is more relevant than older ones for predicting what is going to happen next. It would make more sense to *gradually decrease the weights placed on the older values*. And for the same reason, we should expect forecasts to be less accurate (and therefore to have wider confidence intervals) the farther into the future they are extended. The SMA model does not reflect this. It lacks any underlying theory (a "stochastic equation of motion") to explain why or by how much it should be harder to predict 2 or 3 periods ahead than to predict 1 period ahead, and therefore—very implausibly—its confidence intervals for long-horizon forecasts do not widen at all.

These shortcomings of the SMA problem are addressed by the *simple exponential smoothing model* (SES), which is otherwise known as the *exponentially weighted moving average model,* because it weights the past data in an exponentially decreasing manner, analogous to the discounting of cash flows over time. *The SES model is the most widely used time series model in business applications*, partly because it does a good job of forecasting under a wide range of conditions and partly because computationally it is extremely simple. You easily forecast 10,000 different things in parallel using this model. The latter property isn't quite as important as it once was, given the size and speed of modern computers, but it has contributed to the popularity of this model over the last 50+ years.

There are different ways in which you can write the forecasting equation for the SES model. You don't need to memorize them, but it is at least worth seeing them in order to appreciate the intuitive appeal of this model. One way to write the model is to define a series L that represents the current *level* (i.e., local mean value) of the series as estimated from data up to the present. The value of L at time t computed recursively (i.e., from its own previous value) like this:

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

where $\alpha$ is a "smoothing constant" that is between 0 and 1. Thus, the estimated level at time t is computed by *interpolating between the just-observed value and the previous estimated level*, with weights of $\alpha$ and 1-$\alpha$., respectively. This seems like an intuitively reasonable way to use the latest information to update the estimate of the current level. The series L gets smoother as $\alpha$ approaches zero, because it doesn't change as fast with each new observation of Y. The model assumes that the series has *no trend*, so it predicts zero change in the level from one period to the next. Given this assumption, the forecast for period t+1 is simply the estimated level of the series at time t:

$$\hat{Y}_{t+1} = L_t$$

The only reason for defining the separate series L here is to emphasize that what we are doing is estimating a local mean before turning around and using it as a forecast for the next period. (In more general versions of the model, we will also estimate a *local trend*.) It is equivalent to just say that *the next forecast is computed by interpolating between the last observed value and the forecast that had been made for it:*

Simple Exponential Smoothing, version 1

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t$$

Written in this way, it is clear that the random walk model is an SES model with $\alpha$=1, and the constant-forecast model (of which the mean model is a special case) is an SES model with $\alpha$=0. Hence the SES model is an interpolation between the mean model and the random walk model with respect to the way it responds to new data. As such it might be expected to do better than either of them in situations where the random walk model over-responds and the mean model under-responds, and indeed it does.

There is another equivalent way to write the SES forecasting equation in which the *previous forecast error* plays a role. The error made at period t is the defined as the actual value minus the forecast, i.e.,

$$e_t = Y_t - \hat{Y}_t$$

In terms of this, the SES forecast for period t+1 can be expressed as

Simple Exponential Smoothing, version 2

$$\hat{Y}_{t+1} = \hat{Y}_t + \alpha e_t$$

Thus, *the last forecast is adjusted in the direction of the error it made.* If the error was positive, i.e., if the previous forecast was too low, then the next forecast is adjusted upward by a fraction $\alpha$ of that error. This version provides a nice interpretation of the meaning of "alpha," namely that $\alpha$ *is the fraction of the forecast error that is believed to be due to an unexpected change in the level of the series rather than an unexpected one-time event.* Models with larger values of $\alpha$ assume that what they are seeing are significant changes in the fundamental level of the series from one period to the next. In the limit as $\alpha \rightarrow 1$ (which is the random walk model), *all* of the variation from one period to the next is believed to be due to a change in the fundamental level rather than just a temporary deviation. In the limit as $\alpha \rightarrow 0$ (which is the constant model), the fundamental level of the series is assumed to never change and all of the period-to-period variation is attributed to temporary deviations from it.

Last but not least, the SES forecast can also be written as an *exponentially weighted moving average of all past values,* i.e.,

$$\hat{Y}_{t+1} = \alpha[Y_t + (1\text{-}\alpha)Y_{t\text{-}1} + (1\text{-}\alpha)^2 Y_{t\text{-}2} + (1\text{-}\alpha)^3 Y_{t\text{-}3} + \ldots]$$

Simple Exponential Smoothing, version 3

This is obtained by starting from version 1 of the equation and substituting out the prior forecasts one at a time:

$$\begin{aligned}
\hat{Y}_{t+1} &= \alpha Y_t + (1\text{-}\alpha)\hat{Y}_t \\
&= \alpha Y_t + (1\text{-}\alpha)(\alpha Y_{t\text{-}1} + (1\text{-}\alpha)\hat{Y}_{t\text{-}1}) \\
&= \alpha Y_t + (1\text{-}\alpha)\alpha Y_{t\text{-}1} + (1\text{-}\alpha)^2(\alpha Y_{t\text{-}2} + (1\text{-}\alpha)\hat{Y}_{t\text{-}2}) \\
&= \text{etc. etc.}
\end{aligned}$$

The exponentially-weighted-moving-average form of SES model highlights the difference between it and the simple moving average model: *the SES forecast uses all past values but discounts their weights by a factor of 1-$\alpha$ per period, while the SMA model uses only the last m values and gives them equal weights of 1/m.*

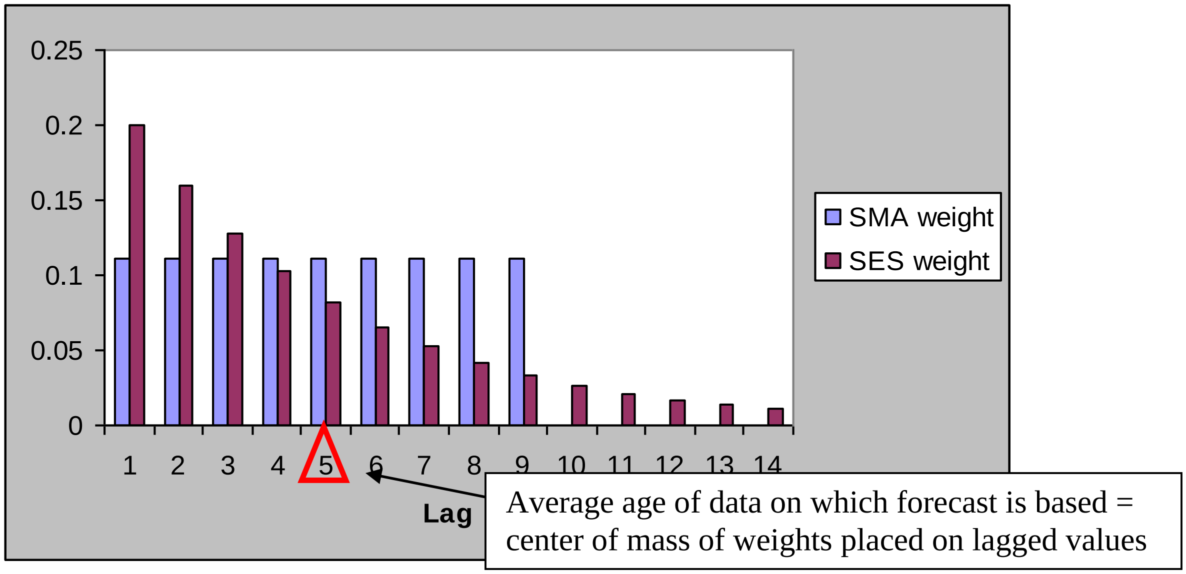
Recall that in discussing the SMA model we introduced the concept of the "average age" of the data in the forecast, which is the amount by which the forecasts of a moving-average model will tend to lag behind turning points, and we saw that it was (m+1)/2 there. The analogous average age of the data in the SES forecast can be obtained by applying the exponential weights to the ages of the previous values of Y that appear in the equation. The value that is 1 period old gets a weight of $\alpha$, the one that is 2 periods old gets a weight of $\alpha(1\text{-}\alpha)$, the one that is 3 periods old gets $\alpha(1\text{-}\alpha)^2$, and so on, which yields:[7]

$$\alpha + 2\alpha(1\text{-}\alpha) + 3\alpha(1\text{-}\alpha)^2 + 4\alpha(1\text{-}\alpha)^3 + \ldots = 1/\alpha$$

Average age of data in SES forecast

This allows a direct comparison between the effect of the smoothing constant $\alpha$ in the SES model and the number of terms m in the SMA model. For a given value of m in the SMA model, if we choose $\alpha$ to satisfy $1/\alpha = (m+1)/2$, then the forecasts of both models will tend to lag behind turning points by exactly the same amount. For example, if we choose m=9 in the SMA model, it yields an average age of 5, which is the same as that of an SES model with $\alpha$=0.2. The weights applied to the lagged values by the two models for these choices of parameters are shown the following figure. For each model the average age is the "center of mass" of the weights, i.e., the point along the x-axis in this chart at which the bars would balance if they were made of a uniform solid material. In this case the center of mass of both sets of bars is exactly 5.

[7] This formula is not supposed to be obvious, but it follows from the fact that the infinite sum $1 + 2x + 3x^2 + 4x^3 + \ldots$ turns out to be equal to $1/(1\text{-}x)^2$. Letting x = 1-$\alpha$ you get the formula.




So, the SES model with α=0.2 behaves similarly to the SMA model with m=5 as far as responsiveness to turning points is concerned. But it does not behave quite the same in other respects: it gives relatively more weight to the 3 most recent values, relatively less weight to the 4th through 9th most recent values, and some positive weight rather than zero weight to the ones before that. Therefore it is a bit more responsive to the most recent events than is the SMA model, which is a good thing. This also means it is more sensitive to the sudden arrival of unusually large or small values than the SMA model, which is perhaps a bad thing if the data distribution is fat-tailed (i.e., contains occasional outliers). But it also has the advantage that data that has already entered the moving average is depreciated thereafter at a constant rate, gradually declining in its impact, whereas the SMA model drops an old value entirely out of the equation when it reaches the end of the moving window. In the 9-term model, the weight of an observation suddenly drops from 1/9 to zero when it is 10 periods old. This sudden dropping of older values introduces some unwanted noise in the SMA model when especially large values drop out of the equation, and on the whole this is probably worse than over-responding to a large value in the most recent period. Overall, the SES model is superior to the SMA model in responding a bit more quickly to the newest data while treating older data more even-handedly, when the models otherwise yield the same average age. Another advantage of the SES model is that its smoothing parameter α is continuously variable over the unit interval and can therefore be optimized to minimize the mean squared error by using tools such as Solver in Excel.

Let's return to our previous example and try the SES model with with α=0.5, α=0.3, α=0.2., and α=0.1. The corresponding average ages are 2, 3.33, 5, and 10, which are what you get with SMA models with m=3, m=5 (approximately), m=9, and m=19. Here is the model comparison table from Statgaphics, including the random walk model as well.

**Models**
(A) Random walk
(B) Simple exponential smoothing with alpha = 0.5
(C) Simple exponential smoothing with alpha = 0.3004
(D) Simple exponential smoothing with alpha = 0.2
(E) Simple exponential smoothing with alpha = 0.1

**Estimation Period**

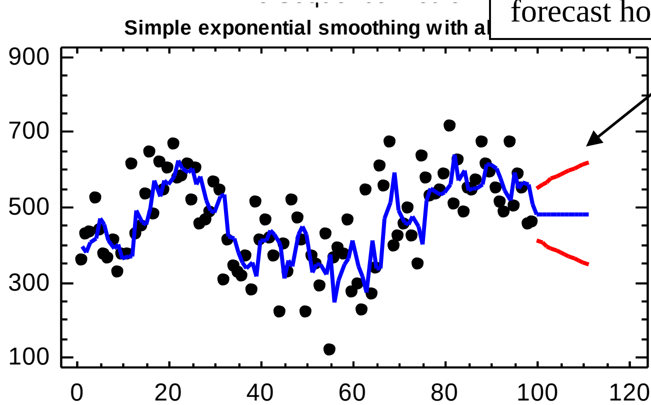
| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 121.759 | 93.2708 | 23.6152 | 1.04531 | -5.21856 |
| (B) | **101.053** | **76.7164** | **14.605** | 1.70418 | -4.94903 |
| (C) | **98.3785** | **75.0555** | **18.9901** | 3.21087 | -4.85317 |
| (D) | **99.5981** | **76.3239** | **12.528** | 5.20827 | -4.7815 |
| (E) | 106.072 | 82.6292 | 20.7756 | 2.4372 | -4.94159 |

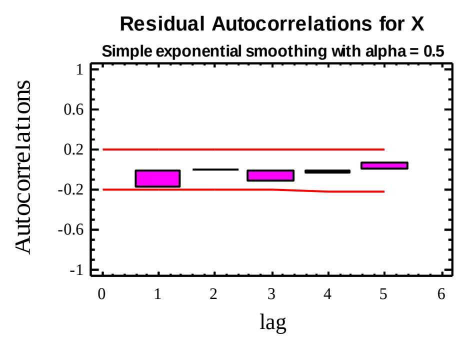
Models B, C, and D (α=0.5, 0.3, 0.2) have about equally good error stats

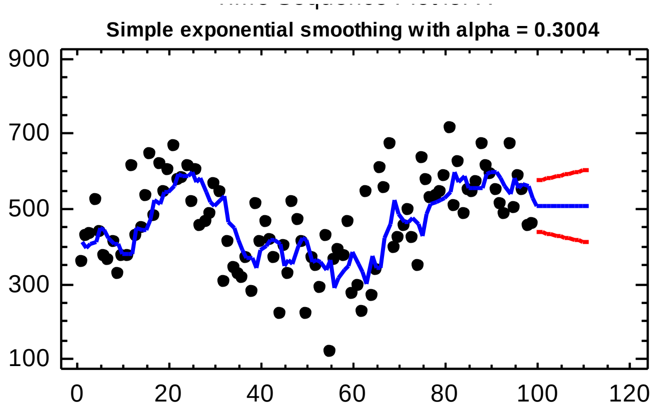
Model A (RW) is clearly worst, models B, C, and D (α = 0.5, .3, .2) are roughly similar in their error stats, and model E (α=0.1) is noticeably worse but better than the RW model. Model C is the one that is obtained by letting Statgraphics optimize α to minimize the RMSE: the optimal value turns out to be **α=0.3004**.[8]

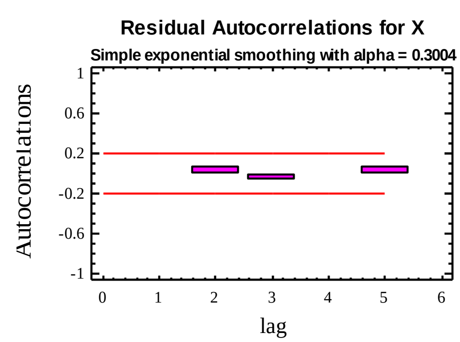
The plots of forecasts and residual autocorrelations of the four SES models are shown below, and they look nearly identical to the corresponding plots for the SMA models. It appears that α=0.1 smooths the data way too heavily, lagging too far behind (namely by 10 periods) and yielding positively autocorrelated errors. The models with smaller α's apply increasing amounts of smoothing and are decreasingly responsive to recent events, and the overall autocorrelation pattern gradually shifts from negative to positive as α decreases (as it does when m increases), although the individual autocorrelations at the first couple of lags in the first three models are technically not significant.

With α=0.5, confidence intervals widen significantly at longer forecast horizons, but not nearly as much as for the RW model.



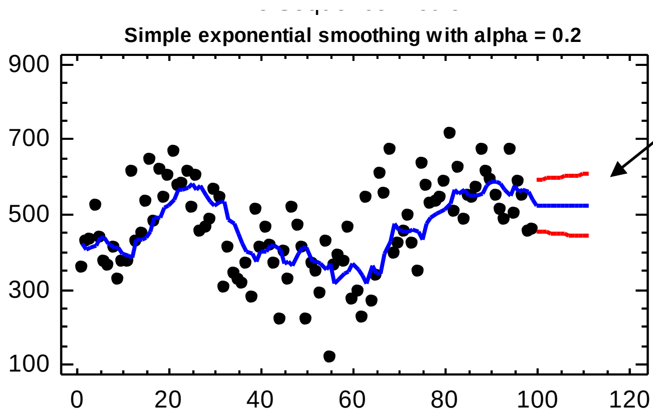



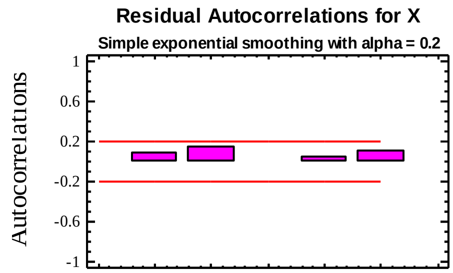



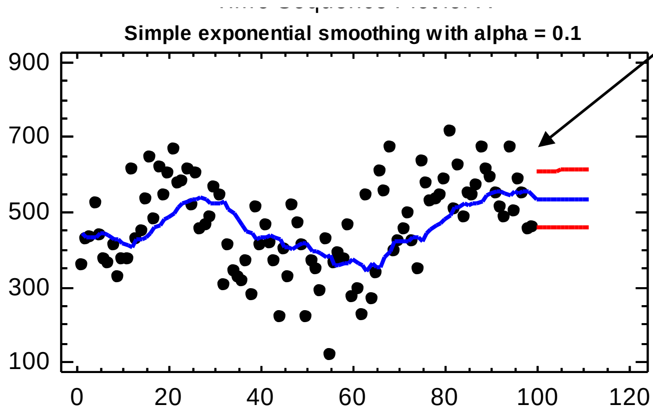

[8] If you re-run the estimation, you may get slight different results in the 3rd or 4th decimal places, but this is not significant. Algorithms for nonlinear optimization don't get "exact" results—they stop when they get "close".

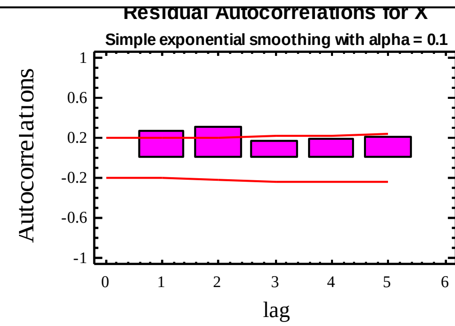
With α=0.2, confidence intervals widen comparatively little at longer horizons.




As α gets smaller, the forecasts get smoother-looking and lag farther behind turning points, as they do when m gets larger in the SMA model.




More importantly, notice that the *confidence intervals grow wider as the forecast horizon lengthens* (unlike those of the SMA models), *and they do so more rapidly for larger values of* α. The confidence intervals for the α=0.5 model and α=0.2 model are about the same for 1-step-ahead forecasts, because the two models have about the same RMSE. However, those of the α=0.5 model grow wider at longer forecast horizons much faster than those of the α=0.2 model. The standard error of the 1-step-ahead forecasts of the SES model is (approximately[9]) the RMSE, and the standard error of the k-step-ahead forecast is obtained by scaling up it as follows:[10]

$$SE_{fcst(k)} = \sqrt{1+(k-1)\alpha^2} \quad SE_{fcst(1)}$$

k-step-ahead forecast standard error for SES model

So the confidence intervals for the k-step ahead forecast of the SES model are wider than those of the 1-step-ahead forecast by a factor of $\sqrt{1+(k-1)\alpha^2}$ compared to a factor of $\sqrt{k}$ for the random walk model. The former factor is smaller than the latter for all α<1, and for any given

[9] If α has been estimated by minimizing mean squared error, then technically we ought to scale up the RMSE by a factor of $\sqrt{n/(n-1)}$ to adjust for the fact that a degree of freedom has been used up, but it wouldn't make a significant difference with a sample of this size.

[10] This equation is not supposed to be obvious, but it follows from the fact that the modeling assumption underlying the SES model is that the random process generating the data is described by the following stochastic equation of motion: $Y_{t+1} = Y_t + \varepsilon_t - (1-\alpha)\,\varepsilon_{t-1}$, where $\{\varepsilon_t\}$ is an i.i.d. sequence of normal random variables.

horizon k it gets smaller with $\alpha$. In the limit as $\alpha \rightarrow 0$, the confidence intervals do not widen at all, as in the case of the mean model.

So, in comparing models B, C, and D ($\alpha$=0.5, 0.3, 0.2 respectively), which have about equally good error stats for 1-step-ahead forecasts, the important differences (so far) are the following:

- ☞ A model with a smaller $\alpha$ yields a smoother (more "resolute") series of forecasts, but lags further behind in following trends and responds later to turning points
- ☞ A model with a larger $\alpha$ is more responsive to recent data and responds more quickly to trends and turning points, but also picks up more "false alarms" about changes in direction.
- ☞ A model with a larger $\alpha$ perceives the distant future to be relatively more uncertain than a model with a smaller $\alpha$, so its confidence intervals widen more rapidly as the forecast horizon lengthens.

The model with the narrowest confidence intervals for longer-term forecasts (model D) is not necessarily the best model just because it has more "self-confidence." And the one that has the absolute lowest RMSE for 1-step-ahead forecasts (model C), is not necessarily best just because it has the absolute lowest RMSE. The amount by which it is better than the other two models on this score is not really significant.

How to choose, then? In a situation like this, probably the best thing to do is to say that there is a range of values of $\alpha$, roughly from 0.2 to 0.5, that appear to be about equally good in terms of the sizes of their typical 1-step-ahead forecast errors,[11] but they make slightly different tradeoffs between smoothness (i.e., not revising the forecast too much from one period to the next) and responsiveness (giving a little more weight to the most recent data). The model that does more smoothing does so because it implicitly assumes that the local mean changes more slowly over time (i.e., that most of the period-to-period variation is merely due to temporary variations rather than changes in fundamentals), in which case the distant future looks more predictable, and data from the distant past looks more relevant for predicting it. For the same reason it lags a bit farther behind in responding to signals that the direction of the wind may have changed. However, the model that is more responsive is also more likely to jump to premature conclusions about changes in fundamentals and it is less confident about predicting the distant future. Perhaps there are other, secondary, criteria that can be used to decide which of these two views of the future is more realistic and how the smoothness-vs.-responsiveness tradeoff should be made. Comparing models on the basis of their *longer-term* forecasting accuracy and on the basis of their *out-of-sample* forecasting accuracy are two other criteria that may be helpful in this regard—we will get to them a bit later.

[11] This result illustrates the "flat minimum" principle in convex optimization: the mean-squared-error function that is being minimized is a quadratic function of the model parameter in the vicinity of the optimum, which is to say that its graph looks like a parabola, which is flat at the bottom. Therefore, minor departures from the optimal parameter values are not really much worse.

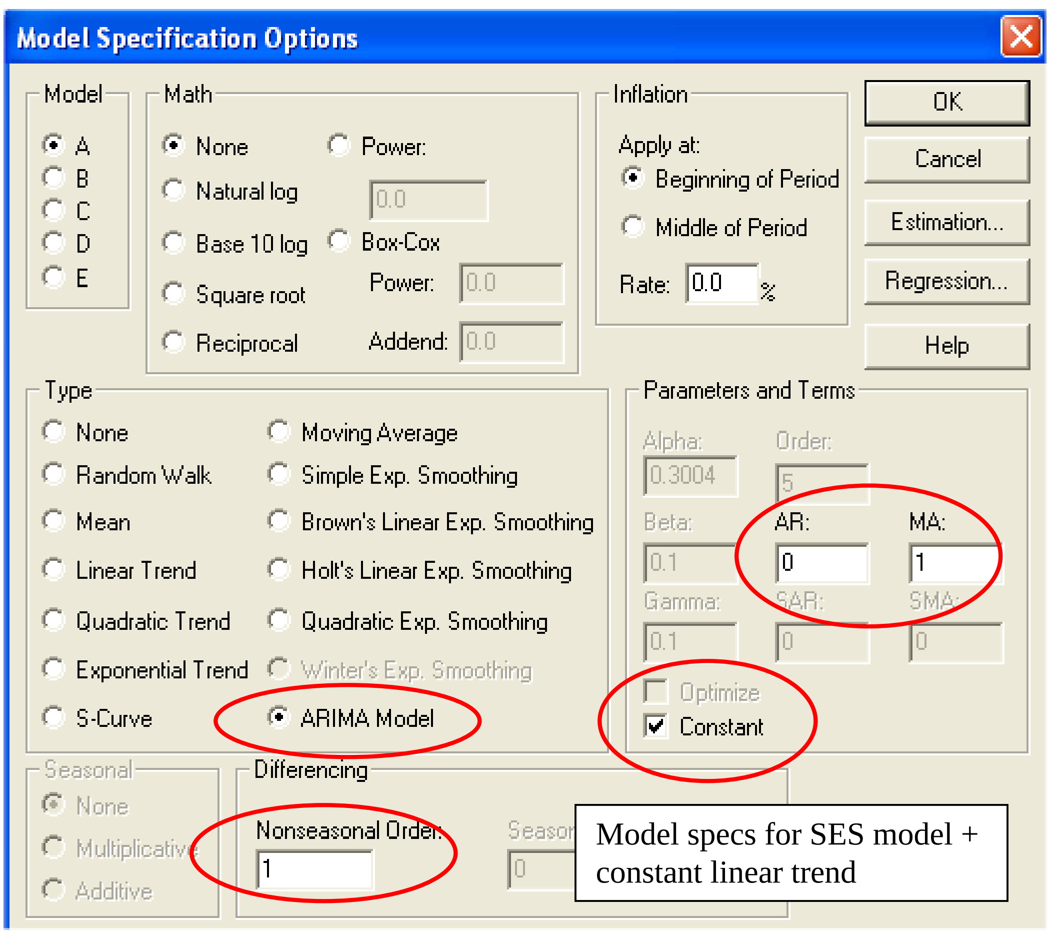
This discussion about how to choose $\alpha$ in an SES model emphasizes the following point, which I will repeat over and over throughout the course: *all models are based on assumptions about how the world works*, and you need to understand what the assumptions are and (ideally) you should believe in the assumptions of your chosen model and be able to explain and defend them. You can't just push a button and let the computer make all your assumptions for you.

What about trends? The SES model, like the SMA model, assumes that there are no trends in the data, either short-term or long-term. However, like the SMA model, it can be modified to incorporate a long-term linear trend by merely adding a drift term to the forecasting equation:

Simple Exponential Smoothing with Trend

$$\hat{Y}_{t+1} = Y_t - (1-\alpha)e_t + d$$

where d is the average long-term growth per period. This model can be directly fitted in Statgraphics as a special case of an ARIMA model: it is a so-called "ARIMA(0,1,1) with constant" model. In the Forecasting procedure, choose the ARIMA model type, and use the following settings: Nonseasonal Order of Differencing = 1, AR=0, MA=1, and constant box checked.




Model specs for SES model + constant linear trend

The model fitting results (in the Analysis Summary report) look like this:

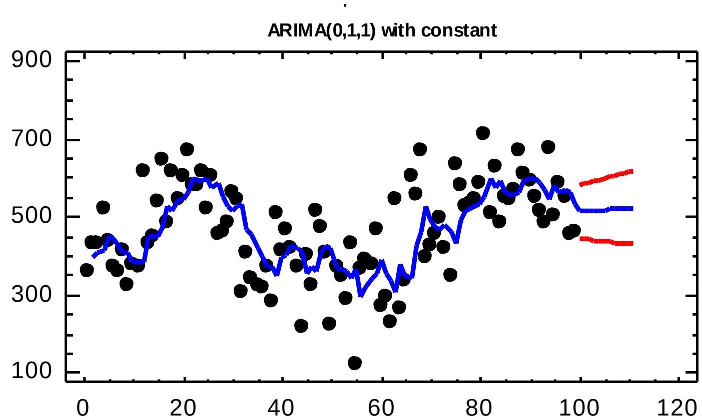
**Forecast Summary**
Nonseasonal differencing of order: 1
Forecast model selected: ARIMA(0,1,1) with constant
Number of forecasts generated: 12
Number of periods withheld for validation: 0

| | *Estimation* | *Validation* |
|---|---|---|
| *Statistic* | *Period* | *Period* |
| RMSE | 99.7012 | |
| MAE | 75.4954 | |
| MAPE | 12.171 | |
| ME | 0.517666 | |
| MPE | -5.53573 | |

**ARIMA Model Summary**

| *Parameter* | *Estimate* | *Stnd. Error* | *t* | *P-value* |
|---|---|---|---|---|
| MA(1) | 0.705504 | 0.07401 | 5.3255 | 0.000000 |
| Mean | 1.03443 | 3.1266 | 0.330849 | 0.741479 |
| Constant | 1.03443 | | | |

The MA(1) parameter in the ARIMA model corresponds to 1-$\alpha$ in the SES model,[12] so here the estimate is $\alpha$ = 1 - 0.7055 = 0.2945, which is about the same as in the SES model without the trend. The Mean parameter in the ARIMA model is the *mean difference between periods*, i.e., the trend. So, d = 1.03443 in the corresponding SES-with-trend equation. The plot of forecasts looks like this:




If you look very closely, you will see that there is indeed a very slight trend. In fact, the average level of the series has risen from roughly 400 to 500 over 100 periods, which is about 1 unit per period, as estimated by the model. However, the estimated trend is not statistically significant (t=0.33, p=0.74), which is reflected in practical terms by the fact that its value ($\approx$1) is very small compared to the 1-step-ahead forecast standard error ($\approx$100).

[12] We will discuss this connection in more detail when we get to ARIMA models.

Of course, the trend factor in the model is more important for *long-term forecasting* than for 1-step-ahead forecasting, so the error stats and significance tests based on 1-step-ahead forecast errors are perhaps not representative of its usefulness. But here it does not appear that a 1-unit-per-period trend would make a difference in long-term forecasting either. So, there is no compelling *statistical* reason to include a constant linear trend in this model. If there are no other compelling reasons either, based on other facts about the situation you may know, then you should not bother to include it. Simpler is better.

### 3.4 LINEAR EXPONENTIAL SMOOTHING[13]

The SMA models and SES models assume that there is no trend of any kind in the data (which is usually OK or at least not-too-bad for 1-step-ahead forecasts), and they can be modified to incorporate a constant linear trend as shown above. What about short-term trends? If a series displays a varying rate of growth or a cyclical pattern and if there is a need to forecast more than 1 period ahead, then estimation of a local trend might also be an issue. The simple exponential smoothing model can be generalized to obtain a *linear exponential smoothing* (LES) model that computes local estimates of both level and trend. The basic logic is the same, but you now have two smoothing constants, one for smoothing the level and one for smoothing the trend. At any time t, the model has an estimate $L_t$ of the local level and an estimate $T_t$ of the local trend. These are computed recursively from the value of Y observed at time t and the previous estimates of the level and trend by two equations. If the estimated level and trend at time t-1 are $L_{t-1}$ and $T_{t-1}$, respectively, then the forecast for $Y_t$ that would have been made at time t-1 is equal to $L_{t-1}+T_{t-1}$. When the actual value is observed, the updated estimate of the level is computed recursively by interpolating between $Y_t$ and its forecast, $L_{t-1}+T_{t-1}$, using weights of $\alpha$ and 1-$\alpha$:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$

(Holt's) LES level-updating equation

The change in the estimated level, namely $L_t - L_{t-1}$, can be interpreted as a noisy measurement of the trend at time t. The updated estimate of the trend is then computed recursively by interpolating between $L_t - L_{t-1}$ and the previous estimate of the trend, $T_{t-1}$, using weights of $\beta$ and 1-$\beta$:

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

LES trend-updating equation

Finally, the forecasts for the near future that are made from time t are obtained by extrapolation of the updated level and trend:

$$\hat{Y}_{t+k} = L_t + kT_t$$

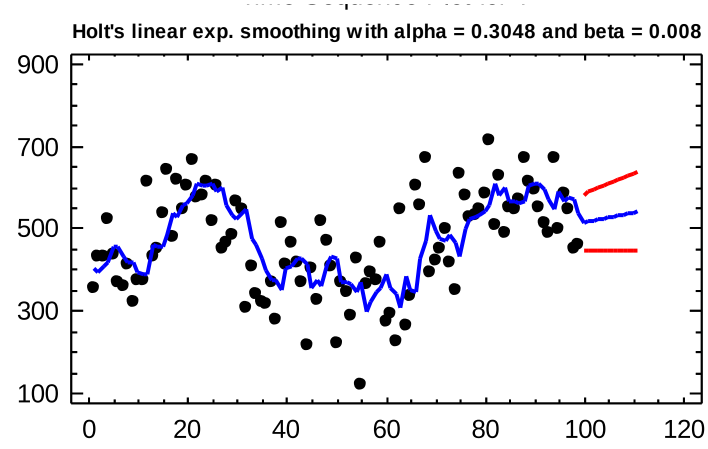
LES forecasting equation

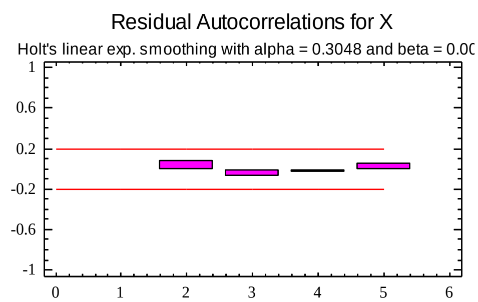
The interpretation of the trend-smoothing constant $\beta$ is analogous to that of the level-smoothing constant $\alpha$. Models with small values of $\beta$ assume that the trend changes only very slowly over time, while models with larger $\beta$ assume that it is changing more rapidly. A model with a large

[13] This is "Holt's" version of the linear exponential smoothing model, which includes separate smoothing constants for level and trend. There is a simpler version—"Brown's" linear exponential smoothing model—which uses a single smoothing constant for both level and trend. In practice, Brown's model tends to yield rather unstable estimates of the trend, because it doesn't smooth the trend heavily enough, and this often leads to excessively wide confidence intervals for longer-term forecasts.

β believes that the distant future is very uncertain, because errors in trend-estimation become quite important when forecasting more than one period ahead.

The smoothing constants α and β can be estimated in the usual way by minimizing the mean squared error of the 1-step-ahead forecasts. When this done in Statgraphics, the estimates turn out to be **α=0.3048** and **β=0.008**. The very small value of β means that the model assumes very little change in the trend from one period to the next, so basically this model is trying to estimate a long-term trend. By analogy with the notion of the average age of the data that is used in estimating the local level of the series, the average age of the data that is used in estimating the local trend is proportional to 1/β, although not exactly equal to it.[14] In this case that turns out to be 1/0.006 = 125. This isn't a very precise number inasmuch as the accuracy of the estimate of β isn't really 3 decimal places, but it is of the same general order of magnitude as the sample size of 100, so this model is averaging over quite a lot of history in estimating the trend. The forecast plot below shows that the LES model estimates a slightly larger local trend at the end of the series than the constant trend estimated in the SES+trend model. Also, the estimated value of α is almost identical to the one obtained by fitting the SES model with or without trend, so this is almost the same model.

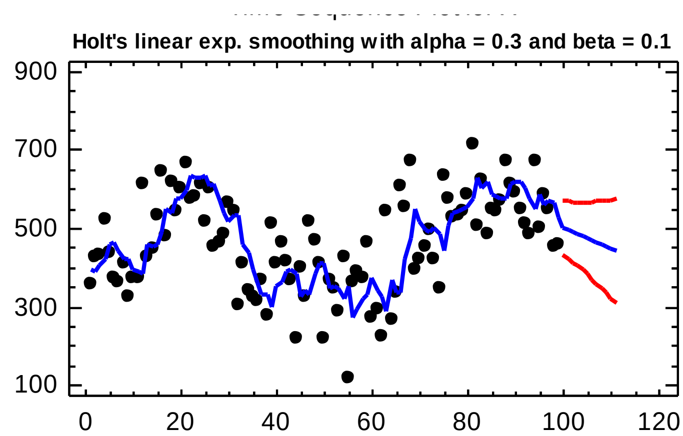


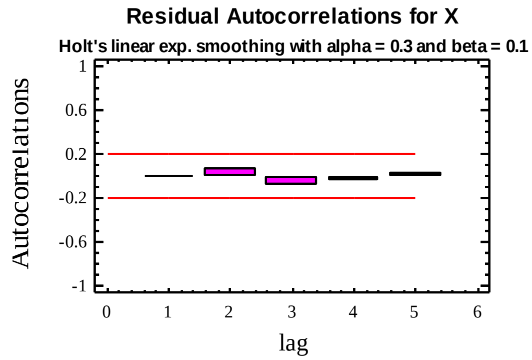




Now, do these look like reasonable forecasts for a model that is supposed to be estimating a local trend? If you "eyeball" this plot, it looks as though the local trend has turned downward at the end of the series! What has happened? The parameters of this model have been estimated by minimizing the squared error of 1-step-ahead forecasts, not longer-term forecasts, in which case the trend doesn't make a lot of difference. If all you are looking at are 1-step-ahead errors, you are not seeing the bigger picture of trends over (say) 10 or 20 periods. In order to get this model more in tune with our eyeball extrapolation of the data, we can manually adjust the trend-smoothing constant so that it uses a shorter baseline for trend estimation. For example, if we choose to set β=0.1, then the average age of the data used in estimating the local trend is 10 periods, which means that we are averaging the trend over that last 20 periods or so. Here's what the forecast plot looks like if we set β=0.1 while keeping α=0.3. This looks intuitively

[14] There is an interaction between α and β in this model in determining how long a baseline is used for level and trend estimation. Usually β should be no greater than 0.1 or the model will get very unstable.

reasonable for this series, although it is probably dangerous to extrapolate this trend any more than 10 periods in the future.




What about the error stats? Here is a model comparison for the two models shown above as well as the three "good" SES models discussed earlier.

**Models**
(A) Holt's linear exp. smoothing with alpha = 0.3048 and beta = 0.008
(B) Holt's linear exp. smoothing with alpha = 0.3 and beta = 0.1
(C) Simple exponential smoothing with alpha = 0.5
(D) Simple exponential smoothing with alpha = 0.3
(E) Simple exponential smoothing with alpha = 0.2

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 98.9302 | 76.3795 | 16.418 | -6.58179 | -7.0742 |
| (B) | 100.863 | 78.3464 | 16.047 | -3.78268 | -5.63482 |
| (C) | 101.053 | 76.7164 | 14.605 | 1.70418 | -4.94903 |
| (D) | 98.3782 | 75.0551 | 18.9899 | 3.21634 | -4.85287 |
| (E) | 99.5981 | 76.3239 | 12.528 | 5.20827 | -4.7815 |

| *Model* | *RMSE* | *RUNS* | *RUNM* | *AUTO* | *MEAN* | *VAR* |
|---|---|---|---|---|---|---|
| (A) | 98.9302 | OK | OK | OK | OK | OK |
| (B) | 100.863 | OK | OK | OK | OK | OK |
| (C) | 101.053 | OK | OK | OK | OK | OK |
| (D) | 98.3782 | OK | OK | OK | OK | OK |
| (E) | 99.5981 | OK | * | OK | OK | OK |

Their stats are nearly identical, so we really can't make the choice on the basis of 1-step-ahead forecast errors within the data sample. We have to fall back on other considerations. If we strongly believe that it makes sense to base the current trend estimate on what has happened over the last 20 periods or so, we can make a case for the LES model with $\alpha$=0.3 and $\beta$=0.1. If we want to be agnostic about whether there is a local trend, then one of the SES models might be easier to explain and would also give more middle-of-the-road forecasts for the next 5 or 10 periods.

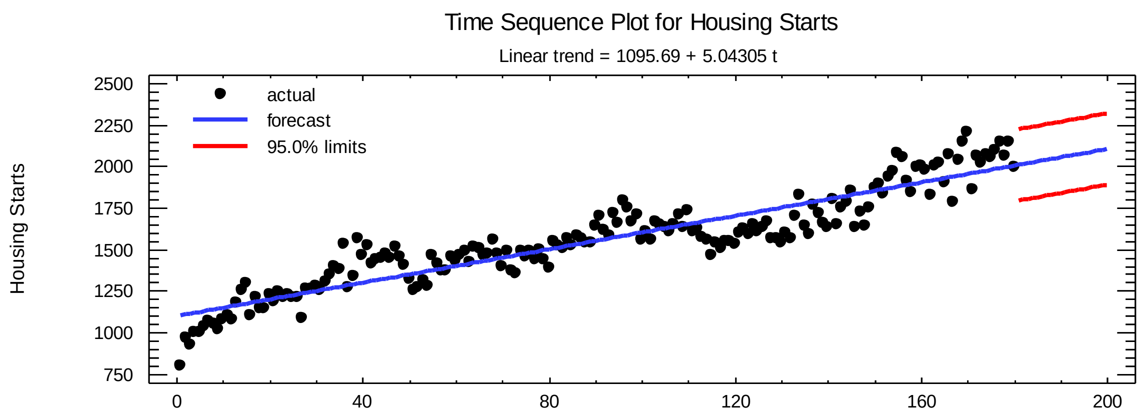
By the way, the second table in the Model Comparison report, which is shown here but was not shown in the earlier comparisons, performs a set of residual diagnostic tests for the presence of any significant "runs" of same-sign errors, autocorrelation, bias (non-zero mean), and changes in variance over time (heteroscedasticity). The key is as follows:

> OK = not significant (p >= 0.05)
> * = marginally significant (0.01 < p <= 0.05)
> ** = significant (0.001 < p <= 0.01)
> *** = highly significant (p <= 0.001)

These are just red flags to alert you to possible problems in case you hadn't already spotted them, and if any significant problem is indicated here, you should look closely at the autocorrelation plots and the plot of residuals versus time and the normal probability plot to see exactly why the red flag went up. Here we do not see any red flags. *These indicators are not your bottom line. You should not throw out a model just because it doesn't get all gold stars here.* Rather, if you see anything funny (particularly some **'s or ***'s), just be sure to look at the relevant plots in order to find out exactly what is going on and whether it indicates that (a) there is a problem with your data, or (b) there is a problem with a model assumption, and/or (c) there is an obvious way in which a model might be improved.[15]

## 5.. A REAL EXAMPLE: HOUSING STARTS REVISITED

To illustrate the application of these models to a more realistic example, consider the history of monthly housing starts series (seasonally adjusted annual rate) from January 1991 to December 2005. If a linear trend model is used, the the fitted values and forecasts look like this in the Forecasting procedure in Statgraphics:

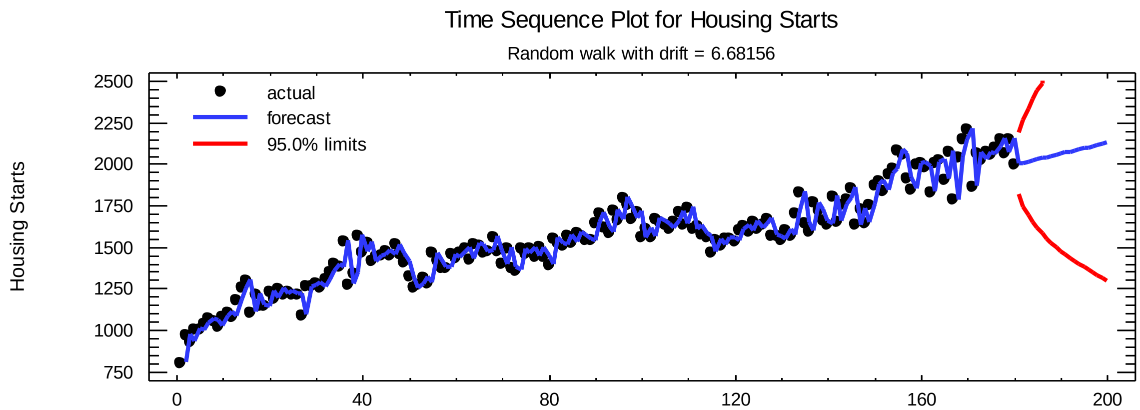



The time index variable here is the row number (1, 2, 3, …) rather than the Date variable used in the Excel regression (1991.000, 1991.083, 1991.167, etc.) , so the model parameters are scaled differently, but it is logically the same model. The estimated trend in this model is 5.043 per month, which is equivalent to the value of 60.517 per year in the previous one.

---

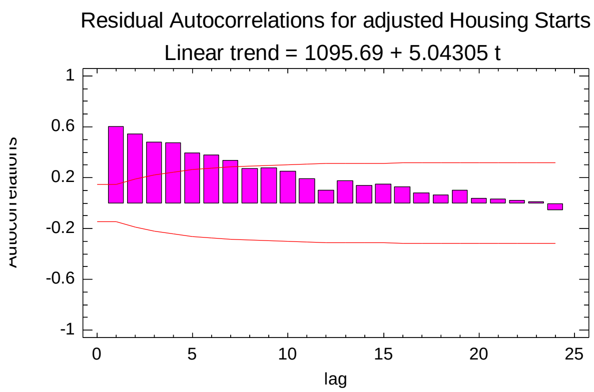
[15] Many years ago I designed this table of red flags in hopes that it would have instructional value and lead to more disciplined modeling, but I've found that many students take it way too seriously and worry about their model getting as many OK's as possible on its report card rather than focusing on the error measures and the story that the data is telling.

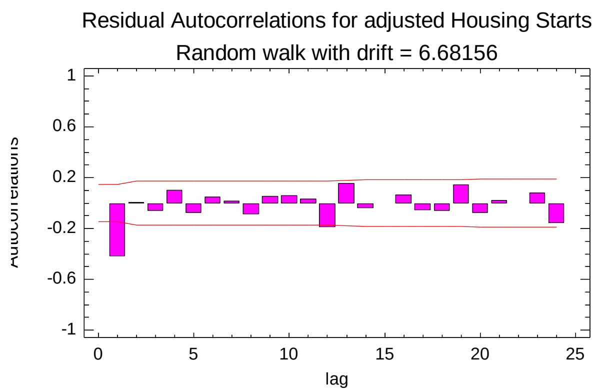
If instead the random walk with drift model is used, the forecasts and confidence limits look like this:




The estimated trend/drift in the random walk model is slightly larger (6.682 per month vs. 5.043 in the linear trend model) because of the difference in estimation methods: the drift term in the random walk model is the slope of a line drawn between the first and last points in the sample, not the slope of a line fitted to the entire sample by least-squares. If you look closely you will see that the random walk model does too much zigging and zagging.

The residual autocorrelation plots of the two models look like this:

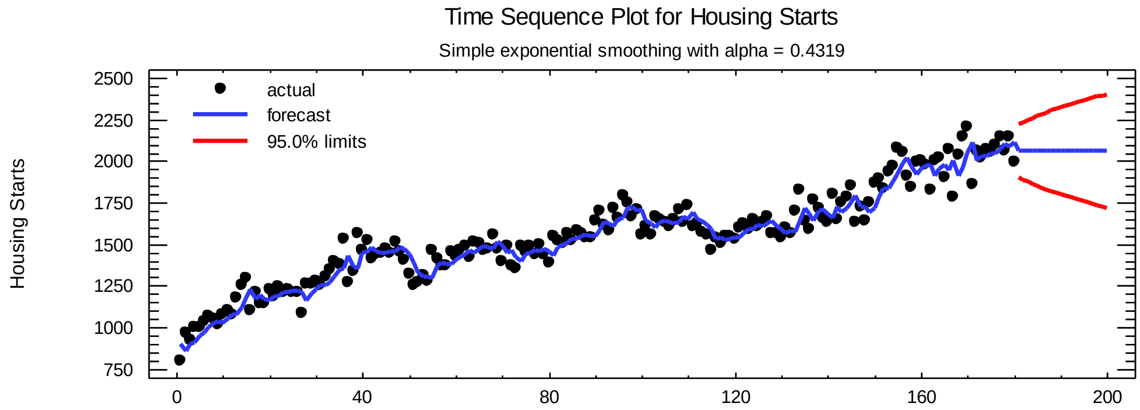


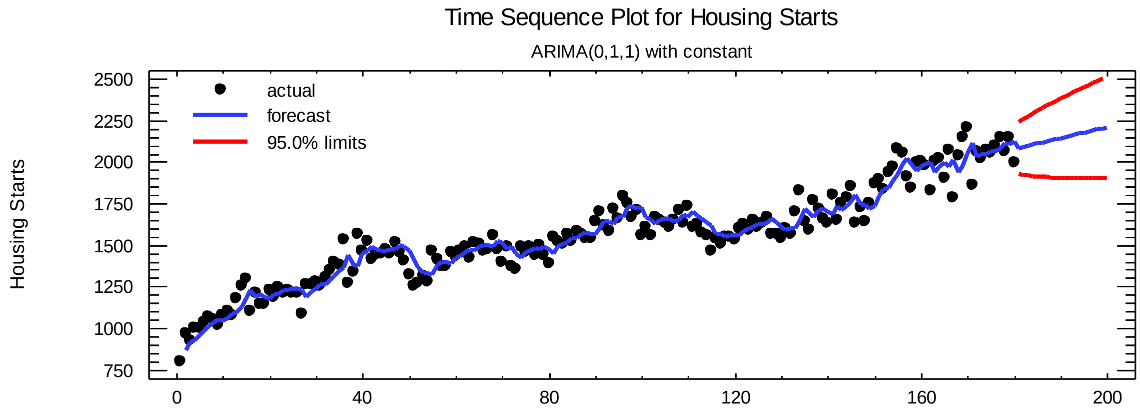

The errors of the linear trend model are seriously positively autocorrelated at lag 1 (and all low-order lags), as we saw earlier, while those of the random walk model are severely *negatively* autocorrelated at lag 1. The RMSE of the random walk model is 94, which is better than the value of 108 obtained with the linear trend model, but judging by the significant autocorrelation in the errors there is room for improvement. Below are the forecast plots that are obtained for the simple exponential smoothing model, ARIMA(0,1,1)+constant model (SES with trend), and Holt's linear exponential smoothing model. The trend in the SES forecasts is zero by assumption, the constant trend estimated by the ARIMA(0,1,1)+constant model (the estimated "mean" in the output) is 6.60, and the local trend estimated by the LES model is 6.64.[16]

[16] The local trend estimated at the end of the sample does not appear explicitly in the output of the LES procedure in

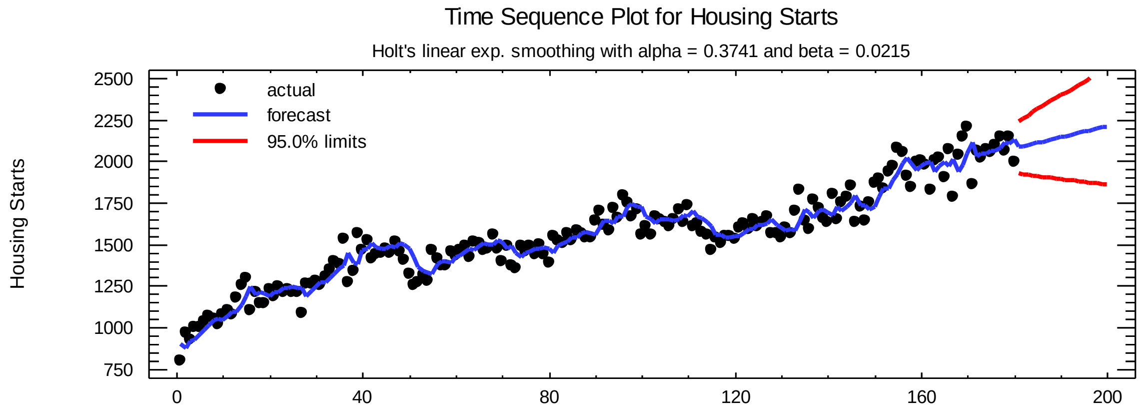
**Forecast plot for SES with no trend:**




**Forecast plot for SES model with constant linear trend (ARIMA(0,1,1)+constant model):**




**Forecast plot for LES model (which has a time-varying local linear trend):**




---

Statgraphics. However, you can determine it by going to the forecast table report and calculating the difference between any two consecutive out-of-sample forecasts. Here the first two out-of-sample forecasts, for periods 181 and 182, are 2084.70 and 2091.34, whose difference is 6.64.

All three of these models have satisfactory-looking residual plots and insignificant residual autocorrelations (not shown here). In fact, their one-step-ahead forecast errors are nearly identical, because they are all doing about the same amount of smoothing of the recent data, and their different approaches to trend estimation don't make much difference for one-period-ahead forecasts. They mainly differ with respect to their forecasts for the more distant future, where the estimated trend makes a difference and where different assumptions about errors in estimating the trend determine the manner in the confidence intervals get wider at longer forecast horizons. The constant trend estimated by the ARIMA(0,1,1)+constant model and the local trend estimated by the LES model are identical in this case, because the LES models is very heavily smoothing its trend estimate. Its estimated value of $\beta$ is only 0.0216, which implies a long baseline for trend estimation.[17] The LES model has slightly wider confidence limits at longer horizons because of its slight uncertainty about how the trend might change over the next 20 periods. Here is the Model Comparison report, which shows that the residual statistics[18] and diagnostics are equally good for the three smoothing models:

**Models**
(A) Random walk with drift = 6.68156
(B) Linear trend = 1095.69 + 5.04305 t
(C) Simple exponential smoothing with alpha = 0.4319
(D) ARIMA(0,1,1) with constant
(E) Holt's linear exp. smoothing with alpha = 0.3702 and beta = 0.0216

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 93.9531 | 72.2278 | 4.65429 | 3.30263E-14 | -0.112802 |
| (B) | 107.78 | 86.7757 | 5.76032 | -1.85689E-13 | -0.539266 |
| **(C)** | **81.8644** | **64.6164** | **4.20202** | **14.9936** | 0.850621 |
| **(D)** | **80.3197** | **62.9115** | **4.08162** | 0.501534 | -0.0767101 |
| **(E)** | **81.101** | **63.6222** | **4.16257** | -0.0432296 | -0.177327 |

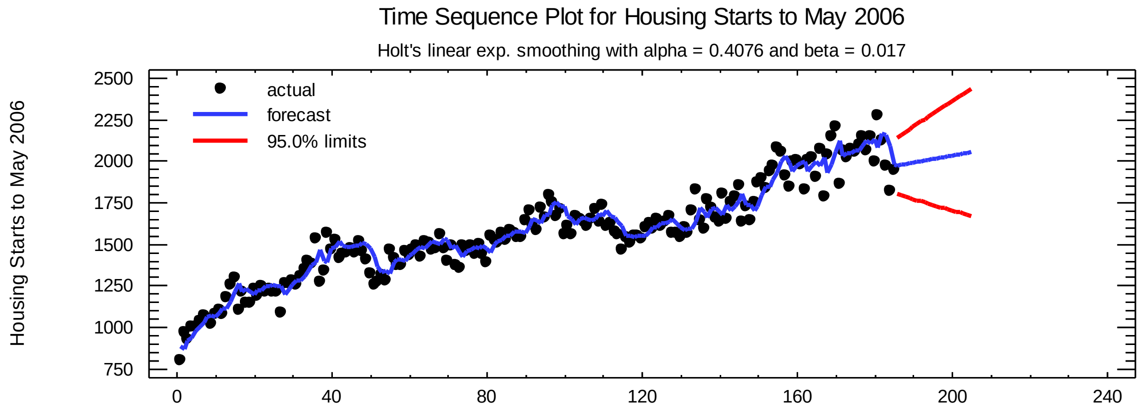
| *Model* | *RMSE* | *RUNS* | *RUNM* | *AUTO* | *MEAN* | *VAR* |
|---|---|---|---|---|---|---|
| (A) | 93.9531 | ** | *** | *** | OK | ** |
| (B) | 107.78 | OK | *** | *** | OK | * |
| (C) | **81.8644** | **OK** | **OK** | **OK** | **OK** | ***** |
| (D) | **80.3197** | **OK** | **OK** | **OK** | **OK** | ***** |
| (E) | **81.101** | **OK** | **OK** | **OK** | **OK** | ***** |

So, there is really not much basis for preferring one of these models over the other as far as 1-step-ahead forecasting over this particular historical period is concerned: they are all averaging the recent data in about the same way. Our choice among the three models should therefore before be based on (i) whether we are really interested in long-term forecasting, and if so, then (ii) whether we wish to assume zero trend or a constant non-zero trend or a time-varying trend. The

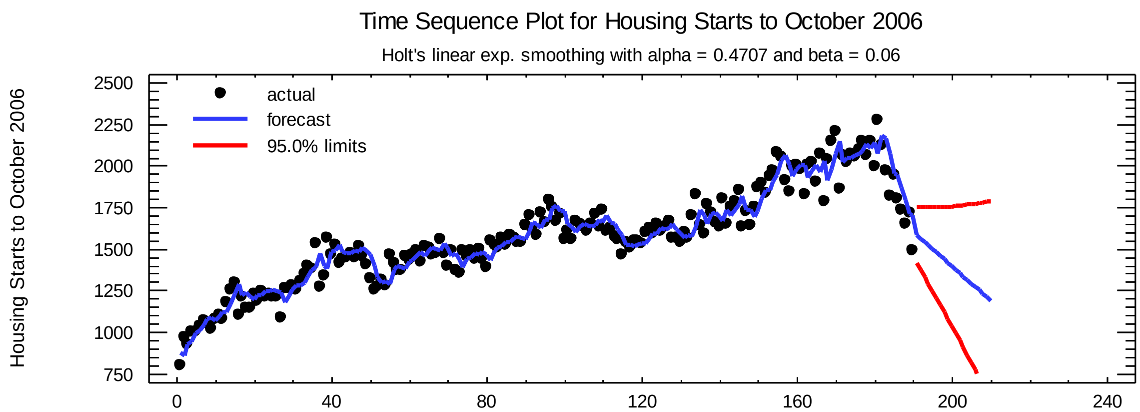
[17] The "average age" of the data used in the trend estimate is not a simple function of $\beta$ in this model, but smaller values mean longer baselines, and 0.02 is very small.

[18] Among the otherwise-similar stats of models C-D-E, one odd thing that stands out is the large value of ME (bias) for model C: it's equal to 15! This is due to the fact that it lacks a drift term, while the data has an upward drift of around 6.7 per period. If the data really has a drift of d per period, then the ME of the SES model will turn out to be on the order of $d/\alpha$, which is 6.7/0.43 = 15.6 in this case. This can be seen from version 2 of the SES equation: the increase in the forecast from periods t to t+1 is equal to $\alpha e_t$, so an average increase of d period results in an amount of bias in which the average value of $\alpha e_t$ is in the general vicinity of d, which implies that the average value of $e_t$ (which is the ME) will tend to be close to $d/\alpha$. This is only an approximate relation because the SES model without drift is unable to exactly reproduce the results of the SES model with drift.

longer-term history of housing starts (going back several more decades) does not show much of a long-term trend—in fact, the trend has turned downward during past recessions—so the SES model would be the conservative choice. However, if we are interested in responding to "turning points" in the business cycle, then we might be interested in the time-varying trend estimation of the LES model. Recall that the original sample extended from January 1991 to December 2005, around the time at which the housing bubble burst and the series sharply reversed direction.[19] Here is the long-term forecast that the LES model would have given if it had been re-run 5 months later, in May 2006:

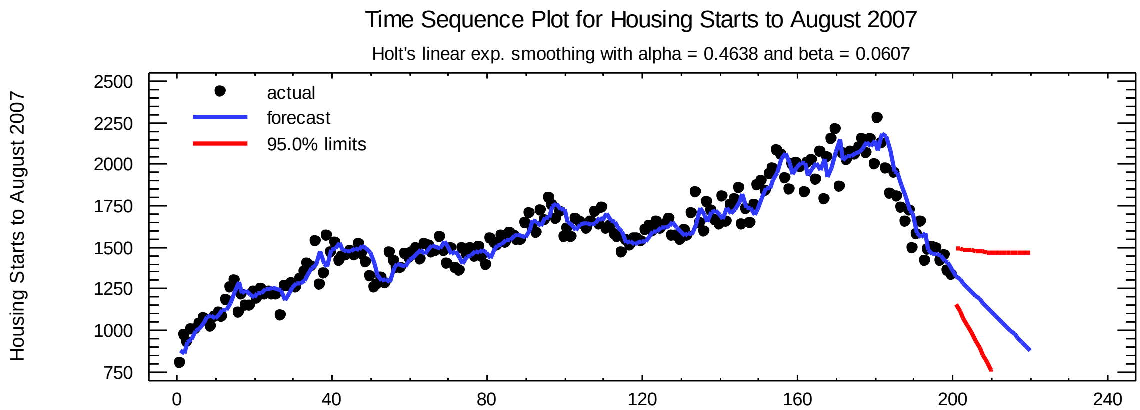



The estimated trend is still positive but has dropped from 6.6 to 4.3 per month, and the estimated value of β is still down around 0.02. Here's how it would have looked after another 5 months, in October 2006:




By then the LES model would have responded to the reversal of direction—the estimated local trend is -20.8 per month and the values of α and β have been revised upward to 0.47 and 0.06, reflecting an adaptation to the fact that the trend is now changing and the level is sinking rapidly. After another 10 months, in August 2007, the forecasts would have looked like this:

[19] The seasonally-adjusted December 2005 value of 1994 was sandwiched in between the peak value of 2273 reached in January and the next-highest value of 2147 that had been reached in November. So, the "dip" in December gave a small hint of the slide that was to begin a few months later.




The estimated local trend is now -23.5, and α and β are about the same as before. This model predicts that the series will fall to 806 in July 2009 (23 months farther into the future), when in fact it fell to 581.

So, the LES model would have reacted to the sharp change of direction within around 10 months, although of course it would not have anticipated it—nor would it anticipate when and where the series would finally bottom out. You ought to look for other evidence to help determine which way the wind is blowing and whether it is starting to change direction. Any smoothing model will lag behind to some extent in responding to unforeseen changes in level or trend.

The bottom line: many time series that arise in business and economics (as well as engineering and the natural sciences) which are inherently non-seasonal or which have been seasonally adjusted display a pattern of *random variations around a local mean value or a local trend line that changes slowly with time*. The first difference of such a series is negatively autocorrelated at lag 1: a positive change tends to be followed by a (smaller) negative one. For time series of this type, a *smoothing or averaging model* is the appropriate forecasting model. The *simple exponential smoothing* model is often a good choice, even though it assumes no trend, and in fact it is probably the most widely used model in applied forecasting. If there is a compelling reason for estimating a non-zero trend for purposes of longer-term forecasting, a constant trend can be implemented via an *ARIMA(0,1,1)+constant* model or a time-varying trend can be implemented via a *linear exponential smoothing* model. However, it is hard to predict changes in trends or the occurrence of turning points merely from the history of the series itself. For such purposes it is important to incorporate information from other sources.

## 6.. OUT-OF-SAMPLE VALIDATION.

In the forecasting examples considered up until now, models have been compared on the basis of their prediction errors within the sample of data to which they were fitted. So, strictly speaking, we have been evaluating the models on the basis of how well they fitted the past, not how well they will predict the future. If we believe that our preferred model is making "correct" assumptions about how the future will look like the past, then its performance in fitting the past data ought to be reasonably representative of its performance in predicting the future. However, this is something that ought to be tested. It can be done by *holding out some of the data* while estimating parameters of alternative models, then freezing those parameter estimates and using them to make predictions for the hold-out data, which is called "out-of-sample validation" or "backtesting." What you hope to find when you do this is that the statistics of the errors of the predictions for the hold-out data look very similar to those of the predictions for the sample data. This is a sign that (a) the general assumptions of the model are valid, and (b) the specific values of the parameters that were estimated from the sample are good ones.

In some cases, though, a model may yield much larger errors in the validation period than in the estimation period because the statistical properties of the data have gradually changed over time, i.e., it is not really statistically stationary. The latter situation is the one you really need to watch out for. A big danger in forecasting is that you may end up with a model that does a great job of predicting the past but not so well in predicting the future. The point of out-of-sample validation is to try to make sure this doesn't happen to you.

When fitting *regression* models to data that are not necessarily time series, it is possible to *randomly select* the set of observations to hold out. However, when working with time series models, the hold-out data usually consists of a *validation period* at the very *end* of the time series. In other words, only some of most recent data is held out. There are two reasons for this. One is that you are most concerned with how well the model is going to perform in the near future, not how well it performed at randomly chosen moments in the past. Another is a purely technical one: by definition, time series models generate their forecasts from prior values of the same data. If you knock holes in the data sample at random locations, this makes it tricky to calculate the forecasts for the values that follow the holes. (Not impossible, just tricky.) In the Forecasting procedure in Statgraphics, the input panel has a box for specifying how many values at the end of the series to hold out for validating the models. The parameters of the 5 selected models will be estimated (only) from the earlier data, and these estimates will then be used to compute forecasts for the validation period. The model comparison table will then show separate sets of comparisons for errors made in the estimation period and in the validation period.

Often you will find that the RMSE's of the models are bigger in the validation period than in the estimation period, and in some cases this is due to greater volatility and/or larger absolute values occurring at the end of the time series. If the data exhibits exponential growth due to compounding or inflation, then it will display greater volatility in absolute terms toward the end of the series, even if the volatility is constant in percentage terms. In situations like this you may want to use a nonlinear transformation such as logging or deflating as part of your model. But *the error statistics reported in the validation table are for errors made in the original units of the data*. The forecasts of all the models are untransformed if necessary before the errors are calculated and measured for use in the model comparison report.[20] This approach is necessary in order to be able to get head-to-head comparisons of models that may have used different nonlinear transformations. But it means that in situations where a nonlinear transformation is really necessary due to exponential growth, you should *expect* the RMSE's to look worse in absolute terms during the validation period. However, the MAPE's should still be comparable, so *it is usually best to look at MAPE's rather than RMSE's when asking whether a given model performed about as well in the validation period as in the estimation period.*

[20] If a transformation such a log transformation has been used as part of the model specification inside the forecasting procedure, then the parameters of the model are estimated by minimized the squared error in transformed units, but the errors are re-calculated in untransformed units for purposes of the model comparison table.

Here are the results of going back to the original housing starts series (180 observations from January 1991 to December 2005) and re-fitting the same models while holding out the last 30 values for validation:

**Model Comparison**
Data variable: Housing Starts
Number of observations = 180
Start index = 1.0
Sampling interval = 1.0
Number of periods withheld for validation: 30

**Models**
(A) Random walk with drift = 7.1745
(B) Linear trend = 1133.69 + 4.35127 t
(C) Simple exponential smoothing with alpha = 0.4832
(D) ARIMA(0,1,1) with constant
(E) Holt's linear exp. smoothing with alpha = 0.4248 and beta = 0.022

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 83.2946 | 65.7421 | 4.52127 | 6.56179E-14 | -0.0990146 |
| (B) | 100.708 | 80.235 | 5.77409 | -5.00222E-14 | -0.619832 |
| **(C)** | **74.1269** | **58.3682** | **4.06086** | **12.4717** | **0.764953** |
| **(D)** | **73.0105** | **57.4137** | **3.97228** | **0.546787** | **-0.0261183** |
| **(E)** | **73.6749** | **58.0385** | **4.04936** | **-1.57981** | **-0.246515** |

| *Model* | *RMSE* | *RUNS* | *RUNM* | *AUTO* | *MEAN* | *VAR* |
|---|---|---|---|---|---|---|
| (A) | 83.2946 | ** | *** | * | OK | OK |
| (B) | 100.708 | OK | *** | *** | OK | OK |
| (C) | 74.1269 | OK | OK | OK | OK | OK |
| (D) | 73.0105 | OK | OK | OK | OK | OK |
| (E) | 73.6749 | OK | OK | OK | OK | OK |

**Validation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 134.717 | 104.555 | 5.32716 | -2.94116 | -0.376723 |
| (B) | 174.004 | 154.525 | 7.54878 | 147.679 | 7.16786 |
| **(C)** | **113.329** | **94.8812** | **4.80061** | **18.5088** | **0.712985** |
| **(D)** | **110.499** | **90.4399** | **4.59364** | **7.70619** | **0.175233** |
| **(E)** | **111.321** | **90.6211** | **4.60692** | **5.7004** | **0.075438** |

Out-of-sample validation with last 30 points of original data set held out: the 3 "good" models are about equally good in the validation period as well, and MAPE's are reasonably consistent between estimation and validation periods.

The three models that were about equally good when fitted to the whole sample (C, D, and E) are also equally good in both the estimation and validation periods: their RMSE's in the estimation period are all around 73-74 and in the validation period they are around 111-113. (Note that the parameter values are slightly different, because a smaller sample of data has been used for estimation.) The MAPE's are reasonably consistent between the estimation and validation periods (around 4% vs. around 4.6-4.8%), which is reassuring.

Now here is a much tougher validation test. In the housing starts analysis, only data up to December 2005 was used in fitting the original models, but post-bubble data to July 2009 (an additional 43 observations) was available. Here are the results of using the last 43 values for validation of the original models. The entire data set (223 observations in total) was used as the input variable, but the last 43 were held out for validation. This is a tougher test because we know the pattern of the data changed sharply in the validation period.

**Model Comparison**
**Data variable: Housing Starts to July 2009**
**Number of observations = 223**
Start index = 1.0
Sampling interval = 1.0
**Number of periods withheld for validation: 43**

**Models**
(A) Random walk with drift = 6.68156
(B) Linear trend = 1095.69 + 5.04305 t
(C) Simple exponential smoothing with alpha = 0.4319
(D) ARIMA(0,1,1) with constant
(E) Holt's linear exp. smoothing with alpha = 0.3702 and beta = 0.0216

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 93.9531 | 72.2278 | 4.65429 | 3.30263E-14 | -0.112802 |
| (B) | 107.78 | 86.7757 | 5.76032 | -1.85689E-13 | -0.539266 |
| (C) | 81.6367 | 64.6164 | 4.20202 | 14.9936 | 0.850621 |
| (D) | 80.3118 | 62.9123 | 4.08199 | 0.549058 | -0.0737948 |
| (E) | 80.6004 | 63.4696 | 4.14962 | -1.62067 | -0.297699 |

| *Model* | *RMSE* | *RUNS* | *RUNM* | *AUTO* | *MEAN* | *VAR* |
|---|---|---|---|---|---|---|
| (A) | 93.9531 | ** | *** | *** | OK | ** |
| (B) | 107.78 | OK | *** | *** | OK | * |
| (C) | 81.6367 | OK | OK | OK | OK | * |
| (D) | 80.3118 | OK | OK | OK | OK | * |
| (E) | 80.6004 | OK | OK | OK | OK | * |

**Validation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | **110.288** | **90.3837** | **8.02598** | **-35.42** | **-3.94162** |
| (B) | 1047.37 | 912.296 | 113.452 | -895.087 | -112.679 |
| (C) | 122.554 | 100.439 | 9.73816 | -80.4446 | -8.04074 |
| (D) | 146.452 | 125.029 | 12.153 | -112.709 | -11.34 |
| (E) | 122.792 | 100.14 | 4.1413 | -74.1742 | -6.54596 |

When the post-bubble data is used as the validation sample, the models all perform much worse in the validation period in terms of both RMSE and MAPE, and surprisingly the random walk model does best in the validation period.

Here the *random walk* model is actually the best in the validation period, evidently because it does the best job of keeping up with the downward slide after January 2006. Even though it is assuming a small upward drift, the fact that it is not smoothing the data means that it doesn't lag as far behind as the other models. The next two best models are the simple and linear exponential smoothing models, and the ARIMA(0,1,1)+constant model doesn't fare as well, because its constant positive trend assumption is now a minus rather than a plus relative to the two other smoothing models. *All of the models yield significantly larger errors in the validation period than in the estimation period, both in absolute and percentage terms.*

It is rather disappointing that the LES model does not outperform the random walk model in the validation period here. We might have hoped it would do the best job of adapting to the sharp downturn in the trend that occurred at the beginning of 2006. Why didn't it do better? The reason is that its *parameter values* were frozen on the values estimated from the data only up to December 2005, and in particular, the trend-smoothing parameter was stuck on the very small value of 0.02 that was estimated from data when the trend was fairly constant rather than the larger value of 0.06 that was estimated from data where the trend was changing. If we fix the parameters of the LES model on values that give more weight to the most recent data (say $\alpha$=0.46 and $\beta$=0.06), the LES model does in fact perform the best in the validation period, beating the random walk model by a small margin. Notice that although this change in the LES model parameters makes a big difference in the validation period, it does not affect the error stats in the estimation period by very much. This reflects the fact that the trend was pretty consistent in the estimation period, so the precise value of $\beta$ is not so important there.

**Model Comparison**
Data variable: Housing Starts to July 2009
Number of observations = 223
Start index = 1.0
Sampling interval = 1.0
Number of periods withheld for validation: 43

**Models**
(A) Random walk with drift = 6.68156
(B) Linear trend = 1095.69 + 5.04305 t
(C) Simple exponential smoothing with alpha = 0.4319
(D) ARIMA(0,1,1) with constant
**(E) Holt's linear exp. smoothing with alpha = 0.46 and beta = 0.06**

**Estimation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 93.9531 | 72.2278 | 4.65429 | 3.30263E-14 | -0.112802 |
| (B) | 107.78 | 86.7757 | 5.76032 | -1.85689E-13 | -0.539266 |
| (C) | 81.8644 | 64.6164 | 4.20202 | 14.9936 | 0.850621 |
| (D) | 80.3118 | 62.9123 | 4.08199 | 0.549058 | -0.0737948 |
| (E) | 81.8395 | 63.6999 | 4.14049 | -2.14617 | -0.342975 |

| *Model* | *RMSE* | *RUNS* | *RUNM* | *AUTO* | *MEAN* | *VAR* |
|---|---|---|---|---|---|---|
| (A) | 93.9531 | ** | *** | *** | OK | ** |
| (B) | 107.78 | OK | *** | *** | OK | * |
| (C) | 81.8644 | OK | OK | OK | OK | * |
| (D) | 80.3118 | OK | OK | OK | OK | * |
| (E) | 81.8395 | OK | OK | OK | OK | * |

When the trend-smoothing parameter of the LES model is set to a larger value (for more sensitivity to changing trends), it now does best in the validation period.

**Validation Period**

| *Model* | *RMSE* | *MAE* | *MAPE* | *ME* | *MPE* |
|---|---|---|---|---|---|
| (A) | 110.288 | 90.3837 | 8.02598 | -35.42 | -3.94162 |
| (B) | 1047.37 | 912.296 | 113.452 | -895.087 | -112.679 |
| (C) | 122.554 | 100.439 | 9.73816 | -80.4446 | -8.04074 |
| (D) | 146.452 | 125.029 | 12.153 | -112.709 | -11.34 |
| **(E)** | **103.317** | **80.5242** | **7.62903** | **-28.9394** | **-1.79138** |

Bottom line: it is important to do out-of-sample validation of forecasting models as a reality check on whether they can predict future data as well as they fitted past data. What you hope to find is that the model that looks the best based on its performance in the estimation period (and other criteria such as intuitive reasonableness and simplicity and industry acceptance) is also the one that performs the best in the validation period. You also hope to find that its MAPE in the estimation period is reasonably representative of its MAPE in the validation period. Sometimes this is too much to hope for if you know that there has been a recent change in the pattern of the data, and in that case it is more-than-usually important to look for other information outside your data set that can shed light on how you should predict the future.