Chapter 7
Exponential Smoothing Models

1 Introduction

As alluded to in chapters 5 and 6, the benchmark forecast models that we have covered so far do not always work well with economic or business data. For example, while the intercept model will forecast well if the mean of the series never changes, the mean of many economic series changes over time. While 4% RGDP growth was common during economic expansions in the 1950's through 1980's, demographic changes such as an aging population and women having already largely entered the workforce have meant that in more recent expansions, GDP growth has been closer to 2%.

In this chapter we will cover a class of models called **Exponential Smoothing** models. These exponential smoothers provide a flexible way to model time-series data. For example, if you wish to forecast employment at your company using only the historical values, exponential smoothing methods can be very useful. In general, using exponential smoothers will force us to estimate more parameters than our benchmark forecasting models did. However, these additional parameters allow the exponential smoothers to match a much larger array of time-series data then the simple methods. For example, we will see soon that two special cases of Simple Exponential Smoothing (SES) are the intercept model and the random walk model. In other words, SES can forecast like the intercept model, the random walk model, or something in-between.

The models we will consider in this chapter are:

1. Simple Exponential Smoothing (SES)
2. Holt's Exponential Smoothing (Holt's)
3. Holt-Winters' Exponential Smoothing (HW)

For each method, we will introduce two sets of equations: "Component Form" and "Error Correction Form." Both are equivalent and will produce the same forecasts, but depending on

exactly what we are trying to do, it may be more convenient to work with one rather than the other. For example, when we talk about forming the h-period ahead forecast, it will be easier to use the error correction form.

Finally, note that when we introduce these three exponential smoothers, we will do so with "additive" models. However, exponential smoothers can also be written in "multiplicative" terms. In "additive" exponential smoothing models, all estimates are in terms of the units your data series is measured in. In "multiplicative" exponential smoothing models, all estimates are in percentages. For example, if your data is measured in dollars, an additive model might say that your series is growing by \$20 per period, while a multiplicative model might say that your series is growing by 1% per period. We only cover the math for the additive models, but once you understand additive models it is easy to understand the multiplicative models.

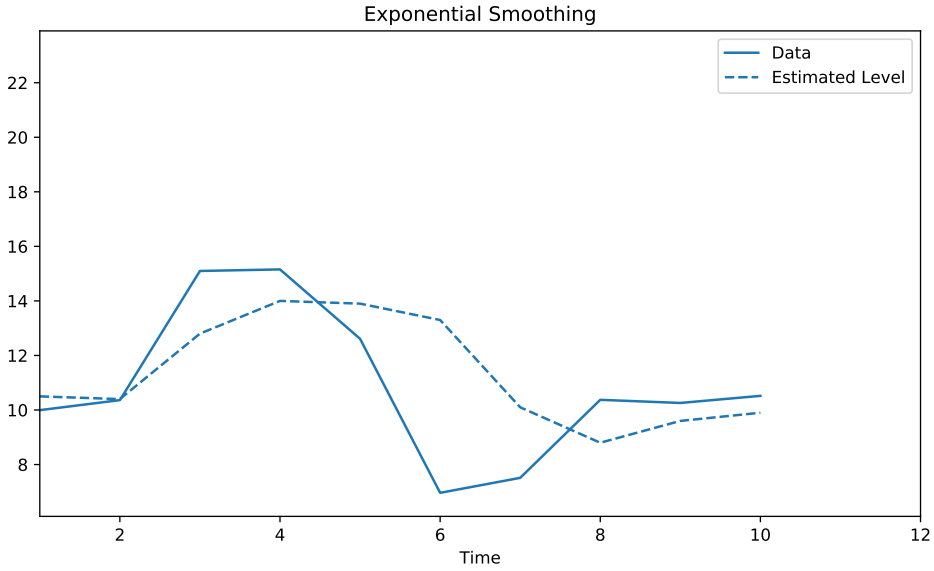
2 Simple Exponential Smoothing (SES)

Simple exponential smoothing performs best when your data series has a potentially changing level, but does NOT have either trend or seasonality.[1] That is, the series does not display any seasonality or long-run trends, but may exhibit short to medium-term deviations from a constant intercept, or even breaks in the intercept. In this case, our goal is to estimate the unobserved "level" of the series using the observed data, and then use this estimated level to form forecasts.

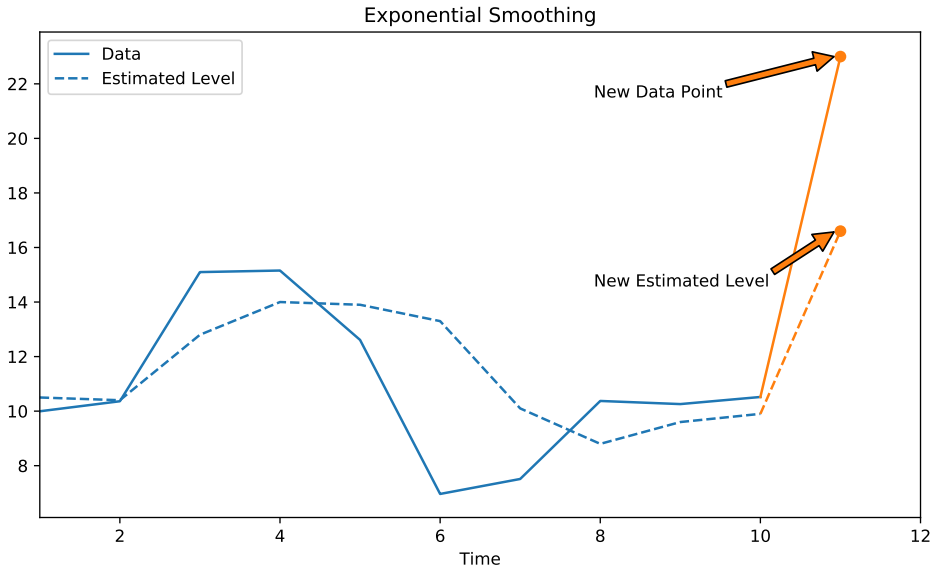
Before describing the model mathematically, let's think about what is going on intuitively. Suppose we have some time-series data, Y, and we would like to form a forecast. Before we forecast, we want to fit a model to the data that we have. When we use SES, we are trying to find the underlying "level" of the data. This "level" is essentially a weighted average of historical values of the time series.

Suppose we estimate the SES model and form in-sample estimates of the level at each point in time. Then, we could make a time plot like the one below, that contains both the data and the estimated level in each time period:

[1] In this chapter "level" is used to refer to the true underlying intercept (or weighted mean) of the series. Mean, intercept, and level will be used synonymously.



Now suppose that we observe a new data point. After we observe this data point, we will update our estimate of the level. If the value of the new data point is larger than the previous value of the level, we will increase our estimate of the level. If the value of the new data point is lower than the previous value, we will decrease our estimate of the level. I've plotted an example below:



To form the estimate of the level, the SES model uses a smoothing parameter, α, that will determine how much we should change our estimate of the level as new data is observed. Small values of the smoothing parameter indicate that we should not change our estimate of the level very much even if we observe a value far away from the level (i.e. our estimate of the level

should be very smooth over time). Large values of the smoothing parameters indicate that if we observe a value far away from the level we should update our estimate of the level by a lot (i.e. our estimate of the level will be less smooth over time).

Mathematically, the SES model is written as two main equations, an "observation" equation and a "level" equation. First let's look at the model in "component form":

$$\begin{aligned} \text{Observation Equation: } & y_t = \ell_{t-1} + \varepsilon_t \\ \text{Level Equation: } & \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1} \\ \text{Error: } & \varepsilon_t \sim N(0, \sigma) \end{aligned}$$

where $0 < \alpha < 1$. In words, this model says that the data at time t, y_t, depends on the level from the previous period, ℓ_{t-1} plus random error. Then, after the data is observed, the level is adjusted. The new value of the level is a weighted average between the new observation, y_t, and the previous level, ℓ_{t-1}, where the weight on each is determined by the smoothing parameter.

2.1 SES as a Weighted Average of Past Data

Using the component form above, starting at the first time period, we can write the model so that the estimated level is given by a weighted average of prior data and our initial guess of the level, ℓ_0. Let's do this for a few time periods to show that the level is simply a weighted average of all past data.

For time period $t = 1$, we have:

$$\begin{aligned} \ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} \\ \ell_1 &= \alpha y_1 + (1-\alpha)\ell_0 \end{aligned}$$

which is a weighted average of the first observation and the initial guess of the level. For time period $t = 2$, we have:

$$\begin{aligned} \ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} \\ \ell_2 &= \alpha y_2 + (1-\alpha)\ell_1 \\ \ell_2 &= \alpha y_2 + (1-\alpha)[\alpha y_1 + (1-\alpha)\ell_0] \\ \ell_2 &= \alpha y_2 + (1-\alpha)\alpha y_1 + (1-\alpha)^2\ell_0 \end{aligned}$$

which is a weighted average of the first two observations and the initial guess of the level. For time period $t = 3$, we have:

$$\begin{aligned}
\ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} \\
\ell_3 &= \alpha y_3 + (1-\alpha)\ell_2 \\
\ell_3 &= \alpha y_3 + (1-\alpha)[\alpha y_2 + (1-\alpha)\alpha y_1 + (1-\alpha)^2\ell_0] \\
\ell_3 &= \alpha y_3 + (1-\alpha)\alpha y_2 + (1-\alpha)^2\alpha y_1 + (1-\alpha)^3\ell_0
\end{aligned}$$

which is a weighted average of the first three observations and the initial guess of the level. We could keep going for any arbitrary value of t.

To make things a bit more concrete, let's use actual values for the smoothing paratmeter, α, so you can see how this pararmeter impacts the weighted avarege. First, let's suppose that $\alpha = 0.9$. Then:

$$\begin{aligned}
\ell_1 &= \alpha y_1 + (1-\alpha)\ell_0 \\
\ell_1 &= 0.9y_1 + 0.1\ell_0
\end{aligned}$$

and

$$\begin{aligned}
\ell_2 &= \alpha y_2 + (1-\alpha)\alpha y_1 + (1-\alpha)^2\ell_0 \\
\ell_2 &= 0.9y_2 + (0.1)(0.9)y_1 + (0.1)^2\ell_0 \\
\ell_2 &= 0.9y_2 + 0.09y_1 + 0.01\ell_0
\end{aligned}$$

and

$$\begin{aligned}
\ell_3 &= \alpha y_3 + (1-\alpha)\alpha y_2 + (1-\alpha)^2\alpha y_1 + (1-\alpha)^3\ell_0 \\
\ell_3 &= 0.9y_3 + (0.1)(0.9)y_2 + (0.1)^2(0.9)y_1 + (0.1)^3\ell_0 \\
\ell_3 &= 0.9y_3 + 0.09y_2 + 0.009y_1 + 0.001\ell_0
\end{aligned}$$

With the value of the smoothing parameter set to 0.9, we can see that the most recent observation receives 90% of the weight in the weighted average. The second-most-recent observation receives 9%, and the final 1% is spread amongst all older observations (and the initial level).

Already for the level at time period 3, the initial guess of the level, ℓ_0, is receiving only 0.001 (or 0.1%) of the weight in the weighted average. This small weight on the initial guess of the level will only decrease as time progresses to $t = 4$ and beyond.

For a more moderate value of the smoothing parameter, older observations will retain larger weight. For instance, take $\alpha = 0.5$. Then we have:

$$\ell_1 = 0.5y_1 + 0.5\ell_0$$
$$\ell_2 = 0.5y_2 + 0.25y_1 + 0.25\ell_0$$
$$\ell_3 = 0.5y_3 + 0.25y_2 + 0.125y_1 + 0.125\ell_0$$

so at the third period, the initial guess of the level is still receiving substantial weight (12.5%). When $\alpha = 0.5$ the most recent observation receives 50% of the weight, the second-most-recent receives 25% of the weight, and the final 25% is spread amongst the previous time periods and the initial level.

Finally, for a small value of the smoothing parameter, older observations (particularly the initial level) receive a much larger share of the weight. For instance, let's consider $\alpha = 0.05$. Then:

$$\ell_1 = 0.05y_1 + 0.95\ell_0$$
$$\ell_2 = 0.05y_2 + 0.0475y_1 + 0.9025\ell_0$$
$$\ell_3 = 0.05y_3 + 0.0475y_2 + 0.045125y_1 + 0.857375\ell_0$$

In this case, the two most recent observations will account for roughly 10% of the weight, with a large fraction of the remainder of the weight coming from the initial level, at least until we get to much later time periods.

2.2 Error Correction From

Recall the component form introduced above:

$$\text{Observation Equation: } y_t = \ell_{t-1} + \varepsilon_t$$
$$\text{Level Equation: } \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$$
$$\text{Error: } \varepsilon_t \sim N(0, \sigma)$$

While this form can be very useful (for example when expressing the level as a weighted average of past data) sometimes it is more convenient to express this model in its "error correction" form. This form will show how the level is adjusted based on the size of the one-period-ahead forecast error.

To put this model in its error correction form, let's start by expanding and rearranging the level equation from the component form:

$$\begin{aligned}
\ell_t &= \alpha y_t + (1-\alpha)\ell_{t-1} \\
\ell_t &= \alpha y_t + \ell_{t-1} - \alpha\ell_{t-1} \\
\ell_t &= \ell_{t-1} + \alpha y_t - \alpha\ell_{t-1} \\
\ell_t &= \ell_{t-1} + \alpha(y_t - \ell_{t-1})
\end{aligned}$$

This says that the level is equal to it's previous value, plus an adjustment based on how poorly the previous level predicted the actual data. From the observation equation, we can see that the last term, $y_t - \ell_{t-1}$ is the error in this model:

$$\begin{aligned}
y_t &= \ell_{t-1} + \varepsilon_t \\
\varepsilon_t &= y_t - \ell_{t-1}
\end{aligned}$$

Substituting this into the rewritten level equation, we have:

$$\begin{aligned}
\ell_t &= \ell_{t-1} + \alpha(y_t - \ell_{t-1}) \\
\ell_t &= \ell_{t-1} + \alpha\varepsilon_t
\end{aligned}$$

Now we can write the model in "Error Correction" form:

$$\begin{aligned}
\text{Observation Equation: } & y_t = \ell_{t-1} + \varepsilon_t \\
\text{Level Equation: } & \ell_t = \ell_{t-1} + \alpha\varepsilon_t \\
\text{Error: } & \varepsilon_t \sim N(0, \sigma)
\end{aligned}$$

Now we can see that if we have a positive error term at time t, the level will increase, and vice-versa for a negative error term. The intuition here is fairly straightforward. If we observe a

positive error, it means that our model under-predicted what actually happened. Some of this error is likely due to random chance — maybe it was hotter than expected today so ice cream sales were higher than expected today. However, some of this error may reflect a long-lasting change in the series — maybe it was hotter than expected today so ice cream sales were higher than expected today. But if it was hotter than expected today, then maybe we are in a heat wave. In that case, ice cream sales will also be higher tomorrow, so we should adjust the "level" of expected sales upwards.

2.3 Estimation

Regardless of the form of the model, in order to use it, we need a value for the smoothing parameter, α, and an initial value for the level, ℓ_0. Once we have those two values, we could advance one time period at a time, estimating the value of the level using the data, the value of α, and the value of the previous level. At the end, we will have **two estimated parameters**, α and ℓ_0. We will have T model predictions (or fitted values), $\ell_1, \ell_2, \cdots, \ell_T$

Since this is a forecasting course, we will not spend too much time on the details of estimation. However, we will typically consider the case when both α and ℓ_0 are unknown. If we tell our computer software that these values are unknown, it will estimate them for us using a process called *Maximum Likelihood* estimation. Basically, it will try several different values for the smoothing parameter and the initial level, form estimates of the rest of the levels and determine the goodness of the in-sample fit. After trying many different values for the smoothing parameter and the initial level, it will return the estimates that give the best in-sample fit.

2.4 Forecasting

After estimation, we will have point-estimates of α and ℓ_0. Using these estimates, we can start at period $t = 1$ (our first observation), and form our estimate of the level at time period 1:

$$\ell_1 = \alpha y_1 + (1 - \alpha)\ell_0$$

where y_1 is the true observed value in period 1. Once we have this estimated level, we can move on to period 2:

$$\ell_2 = \alpha y_2 + (1 - \alpha)\ell_1$$

We continue in this way until we reach the end of the sample:

$$\ell_T = \alpha y_T + (1-\alpha)\ell_{T-1}$$

When we reach the end of the sample, we will have the estimated "level" of the series at all time periods. In math terms, we will have $\ell_1, \ell_2, \cdots, \ell_T$.

Now suppose we wanted to forecast the first period out-of-sample, $T+1$. Recall that the full model, in error correction form, is given as:

$$\begin{aligned} \text{Observation Equation: } & y_t = \ell_{t-1} + \varepsilon_t \\ \text{Level Equation: } & \ell_t = \ell_{t-1} + \alpha\varepsilon_t \\ \text{Error: } & \varepsilon_t \sim N(0, \sigma) \end{aligned}$$

Therefore, at time period $t = T+1$ the equation for y_{T+1} is:

$$y_{T+1} = \ell_T + \varepsilon_{T+1}$$

Under squared error loss, we can find the optimal forecast by taking the expected value of both sides, conditional on the observed data, Y:

$$E(y_{T+1}|Y) = E(\ell_T|Y) + E(\varepsilon_{T+1}|Y)$$

Since ε_{T+1} is out-of-sample, and since the error terms are uncorrelated by assumption, $E(\varepsilon_{T+1}|Y) = 0$. Therefore we have:

$$E(y_{T+1}|Y) = E(\ell_T|Y)$$

Since we were able to compute ℓ_T with the in-sample data, we know that $E(\ell_T|Y) = \ell_T$. So we have:

$$E(y_{T+1}|Y) = \ell_T$$

$$\hat{y}_{T+1|T} = \ell_T$$

In words, the optimal one-period ahead forecast when using SES and squared error loss is the

last in-sample estimate of the level.

Finally, similar to both the mean and random walk methods, the optimal point-forecast at any arbitrary horizon, h, is the same as the optimal one-period ahead forecast. In mathematics:

$$\hat{y}_{T+h|T} = \ell_T$$

In other words, the forecast for any future period is equal to the last estimated level.

2.4.1 When to Use

The SES method is suprisingly flexible, able to model many different series as long as they do not have a trend or seasonality. To help see this, note two special cases of the SES method. First, as $\alpha \to 0$, the SES method forecast will approach the forecast from the intercept method, since the level will never be adjusted over time. Instead, the estimate of the initial value of the level, ℓ_0, would be the mean of the data series and would not be adjusted over time.

On the other hand, as $\alpha \to 1$, the method approaches the random walk method. Recall that in the component form we write the level equation as:

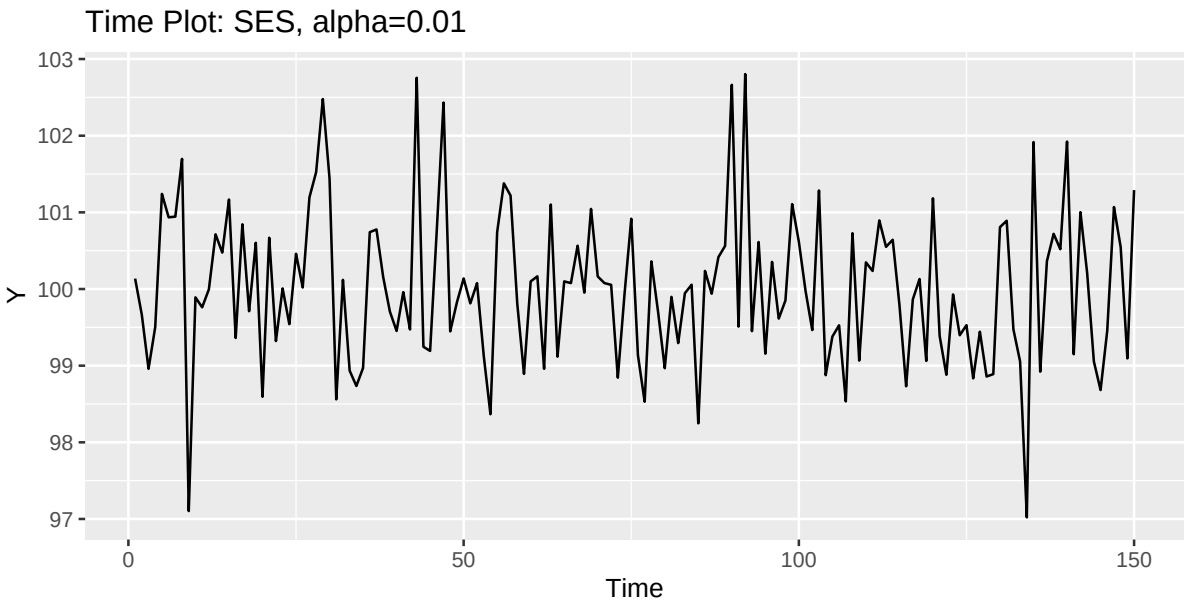
$$\ell_t = \alpha y_t + (1 - \alpha)\ell_{t-1}$$

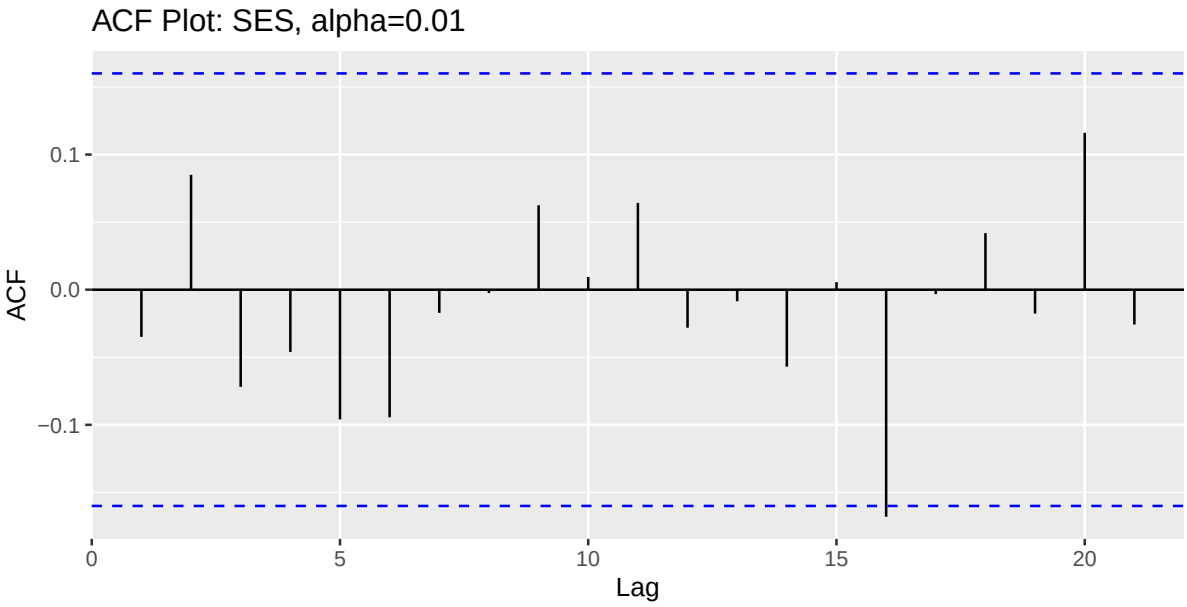
If $\alpha = 1$, then:

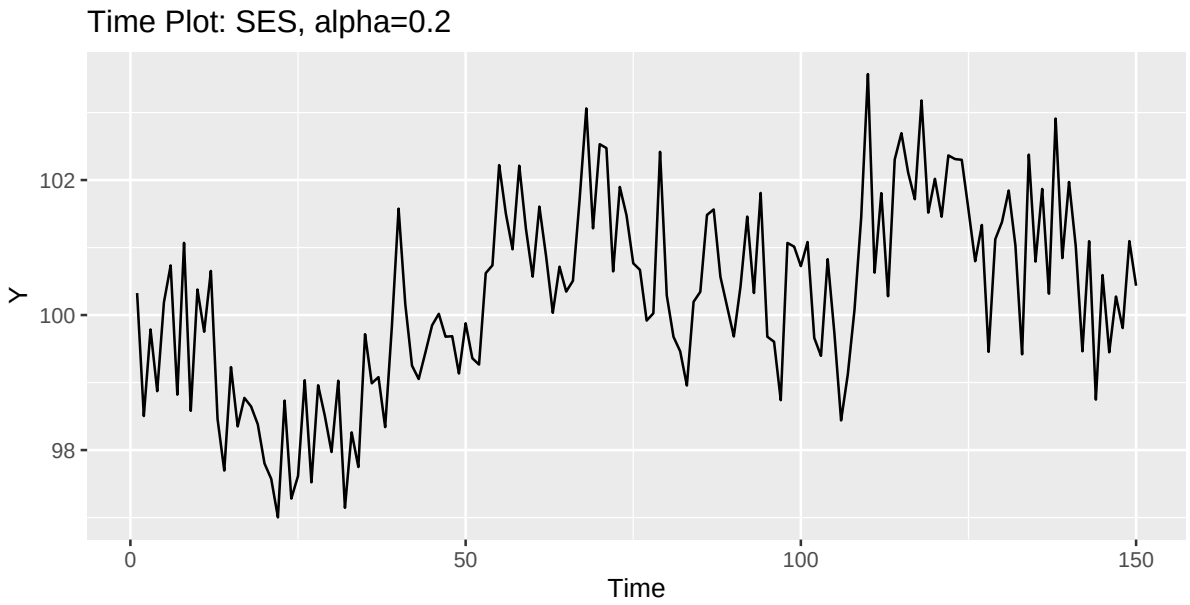
$$\ell_t = y_t + (0)\ell_{t-1}$$
$$\ell_t = y_t$$

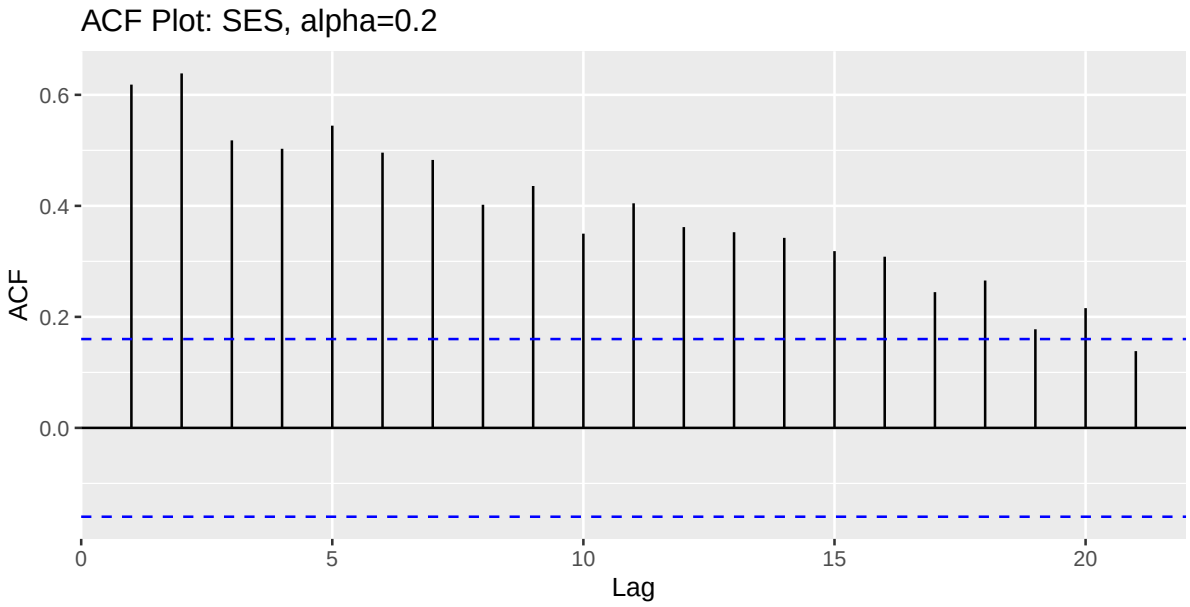
and the estimated level is exactly equal to the last data point. Therefore, the optimal forecast from the SES method would be y_T, the same as it is under the random walk method.

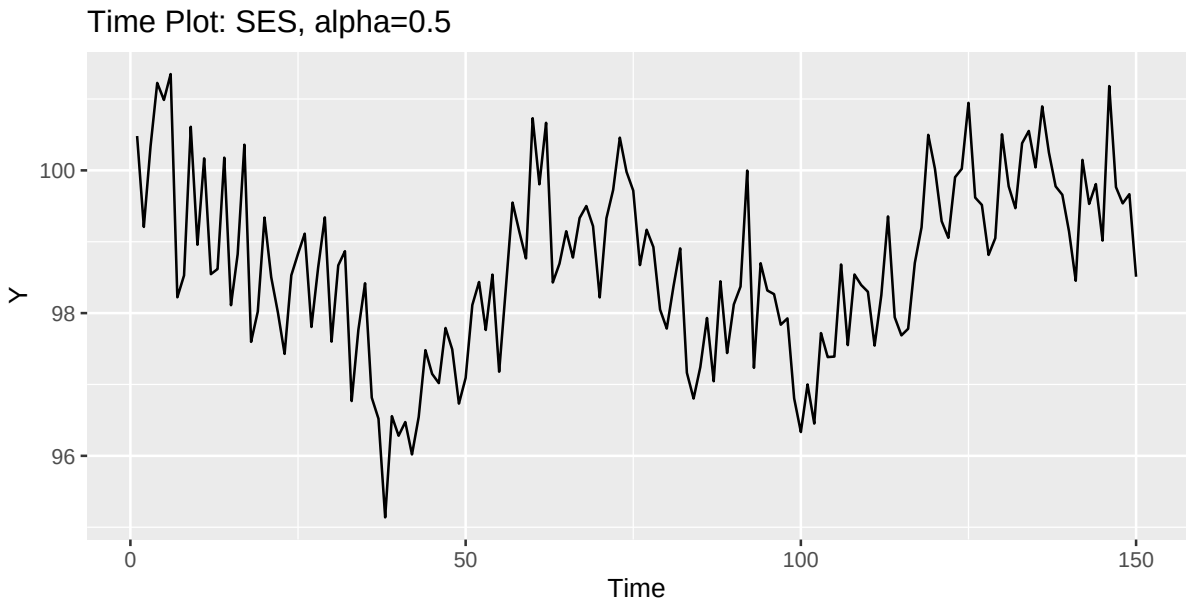
If α is not very close to 0 or 1, then it will allow the level estimate to be adjusted over time, but at a more stable rate then under the random walk method. To get a sense of what type of data the SES method can be used on, I've simulated data with different values of α. If the time plot of your data set looks something like these, you should consider using the SES model to forecast that data.

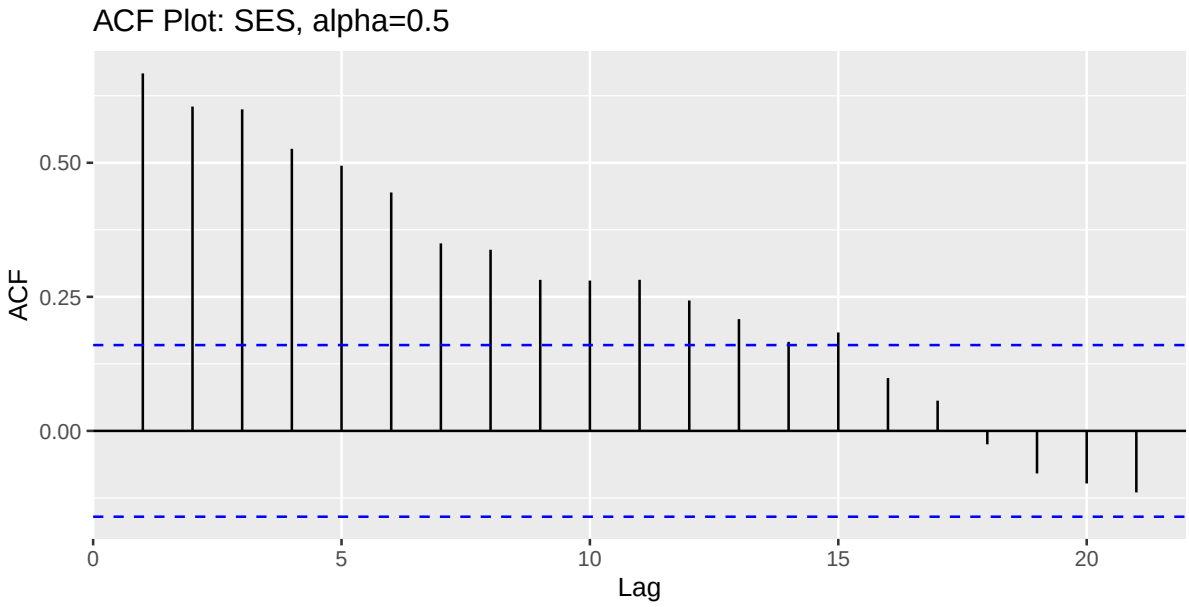

Time Plot: SES, alpha=0.01
103
102
101
100
99
98
97
Y
0
50
100
150
Time

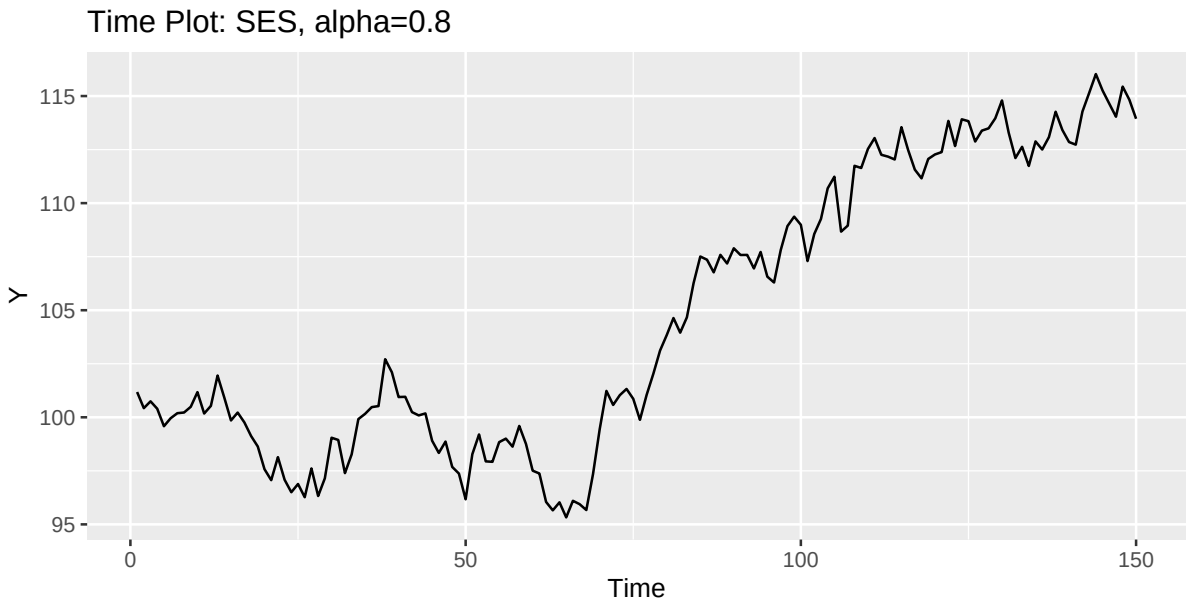

ACF Plot: SES, alpha=0.01
0.1
0.0
−0.1
ACF
0
5
10
15
20
Lag

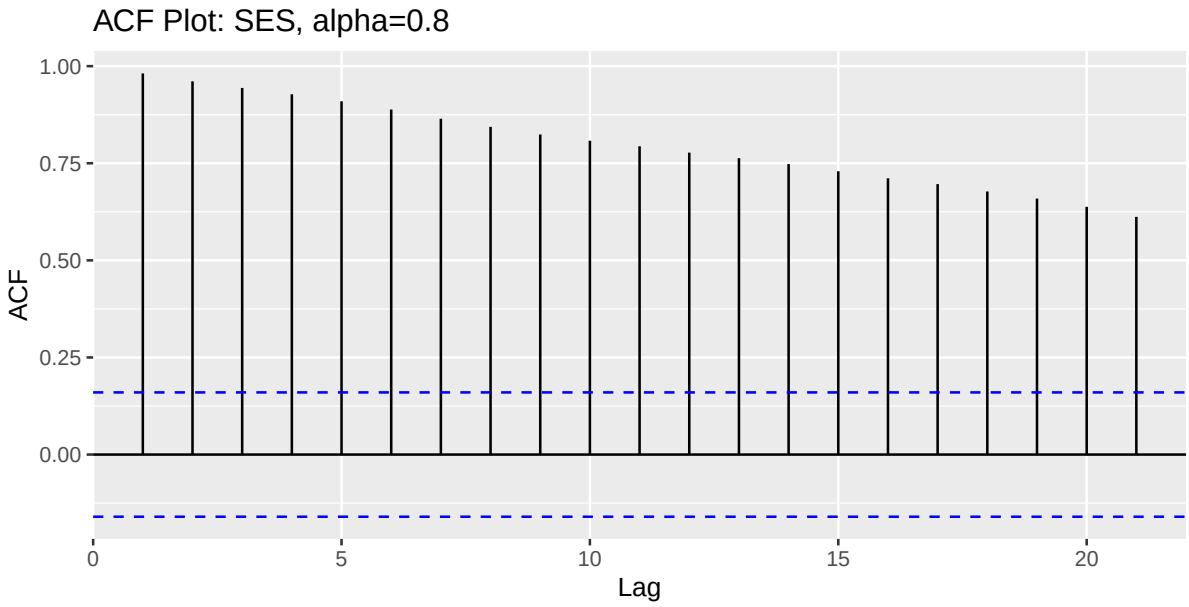

Time Plot: SES, alpha=0.2
Y
102
100
98
0
50
100
150
Time

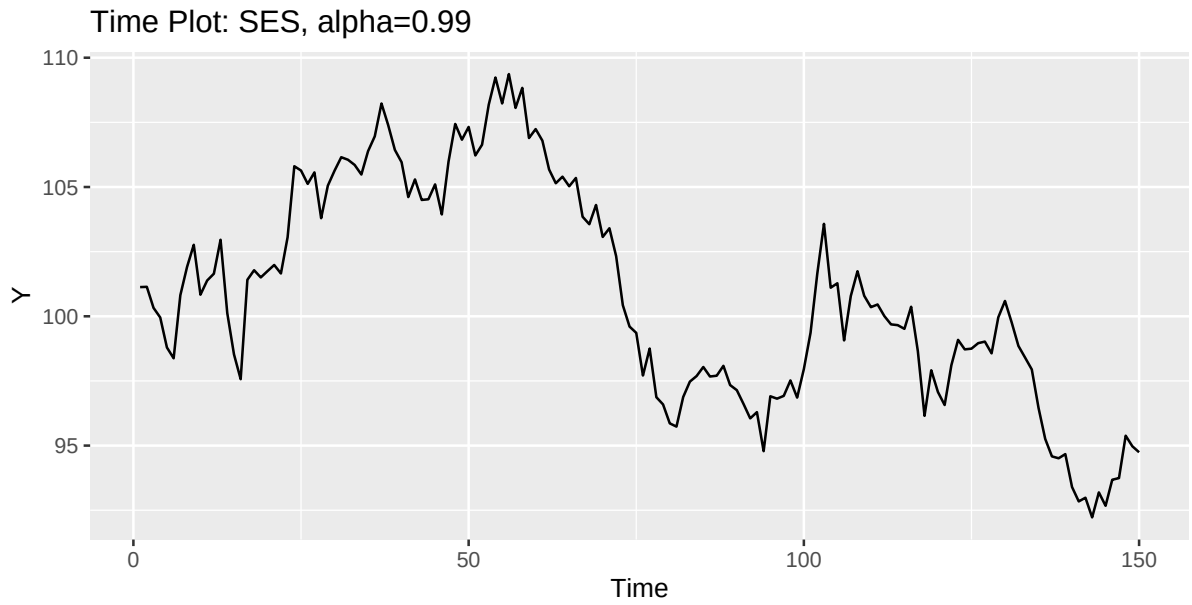

ACF Plot: SES, alpha=0.2
ACF
0.6
0.4
0.2
0.0
0
5
10
15
20
Lag

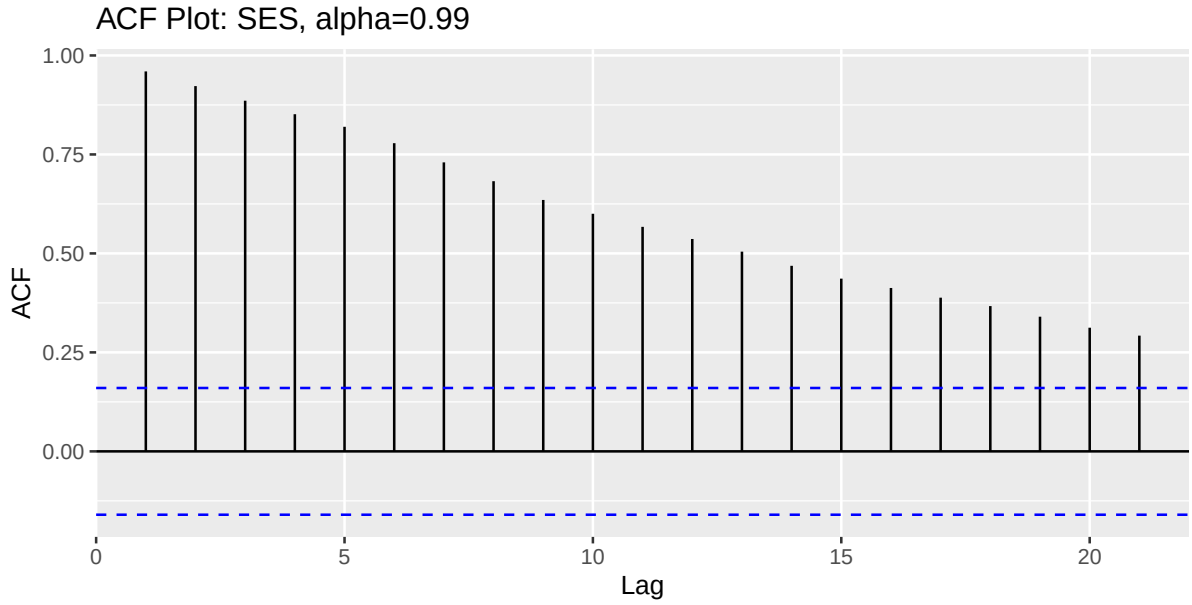

Time Plot: SES, alpha=0.5
Y
100
98
96
0
50
100
150
Time


ACF Plot: SES, alpha=0.5
ACF
0.50
0.25
0.00
0
5
10
15
20
Lag


Time Plot: SES, alpha=0.8
115
110
105
100
95
Y
0
50
100
150
Time


ACF Plot: SES, alpha=0.8
1.00
0.75
0.50
0.25
0.00
ACF
0
5
10
15
20
Lag





Summarizing, the SES model is appropriate to use if your data does not display a trend or seasonality, but an intercept model or random walk model seems inadequate.

3 Holt's Exponential Smoother

In a 1957 paper, Charles Holt extended simple exponential smoothing to allow for a trend. Holt's exponential smoother builds on top of SES to allow for a trend that can change over time. This flexible trend model is very general and can fit a wide variety of data with a trend, including cases when the trend reverses.

Holt's exponential smoother introduces a "trend-smoothing" parameter, β, that acts much

like the α parameter does for the level. Small values of the trend-smoothing parameter (near 0) mean that the trend changes slowly over time, while values of the trend-smoothing parameter near 1 allow the trend to change very quickly over time. Mathematically, Holt's method adds a trend equation to the SES equations. In the "Component" form, it is written as:

$$\begin{aligned}
\text{Observation Equation: } & y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t \\
\text{Level Equation: } & \ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend Equation: } & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1} \\
\text{Error: } & \varepsilon_t \sim N(0, \sigma)
\end{aligned}$$

Where b_t is the estimated linear trend at time t.

As we did with the SES model, let's write this model in error correction form. First note that under this method, using the observation equation, we have:

$$\varepsilon_t = y_t - \ell_{t-1} - b_{t-1}$$

Next, let's use the level equation and rearrange to see if we can make a substitution, similar to what we did in the last section with SES.

$$\begin{aligned}
\ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\ell_t &= \alpha y_t + (\ell_{t-1} + b_{t-1}) - \alpha(\ell_{t-1} + b_{t-1}) \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha(y_t - \ell_{t-1} - b_{t-1}) \\
\ell_t &= \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t
\end{aligned}$$

Let's try the same for the trend equation:

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta) b_{t-1}$$

First, substitute in for ℓ_t using the error correction equation we found above:

$$b_t = \beta(\ell_{t-1} + b_{t-1} + \alpha\varepsilon_t - \ell_{t-1}) + (1-\beta)b_{t-1}$$

$$b_t = \beta(b_{t-1} + \alpha\varepsilon_t) + b_{t-1} - \beta b_{t-1}$$

$$b_t = \beta b_{t-1} + \alpha\beta\varepsilon_t + b_{t-1} - \beta b_{t-1}$$

$$b_t = \alpha\beta\varepsilon_t + b_{t-1}$$

$$b_t = b_{t-1} + \alpha\beta\varepsilon_t$$

So in "Error Correction" form, we have:

$$\text{Observation Equation: } y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$$

$$\text{Level Equation: } \ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$$

$$\text{Trend Equation: } b_t = b_{t-1} + \alpha\beta\varepsilon_t$$

$$\text{Error: } \varepsilon_t \sim N(0, \sigma)$$

The intuition here is similar to the intuition behind the error correction form of SES. If the actual data comes in higher than expected, then some of that is due to random chance, some because we underestimated the level, and some because we underestimated the trend. Therefore, when we observe a positive residual, we increase our estimates of both the level and the trend.

Like in SES, both the error correction & component forms will produce the same estimated trend and level — they are identical, just rearranged, equations. However, for some purposes it may be easier to work with one version rather than the other.

3.1 Estimation

Like the SES model, when we are working with real data, the parameters of Holt's model are unknown and need to be estimated. In addition to the initial condition for the level, ℓ_0, and the level smoothing parameter, α, we need to estimate the initial condition for the trend, b_0, and the trend-smoothing parameter, β. We will let our software package perform the estimation for us. Once estimated, we will treat all **four parameters** as known. Using these parameters, we will be able to form in-sample fitted values for: $\ell_1, \ell_2, \cdots, \ell_T$ and $b_1, b_2, \cdots, b_T$. We can plot these fitted values in time plots to "decompose" the time-series into the level, trend, and error. We will do this in the last section of this chapter.

3.2 Forecasting

If we were interested in forecasting under squared error loss, we could compute the h-period ahead forecast by using the expected value operator. Starting with the one-period ahead forecast, and using the error correction form, we plug in $t = T + 1$ into the observation equation and find:

$$\begin{aligned} y_t &= \ell_{t-1} + b_{t-1} + \varepsilon_t \\ y_{T+1} &= \ell_T + b_T + \varepsilon_{T+1} \\ E(y_{T+1}|Y) &= E(\ell_T|Y) + E(b_T|Y) + E(\varepsilon_{T+1}|Y) \end{aligned}$$

Since the error term is assumed to be iid, mean 0, the last expectation above drops out. Furthermore, since we have estimated the initial conditions and the values of the smoothing parameters, have already have estimates for ℓ_T and b_T that we can plug in. So we have:

$$\begin{aligned} E(y_{T+1}|Y) &= E(\ell_T|Y) + E(b_T|Y) + E(\varepsilon_{T+1}|Y) \\ E(y_{T+1}|Y) &= \ell_T + b_T \\ \hat{y}_{T+1|T} &= \ell_T + b_T \end{aligned}$$

Under Holt's method, as long as $b_T \neq 0$, the h-period ahead forecast will NOT be the same as the one-period ahead forecast. However, there will be a pattern. Finding the optimal two-period ahead forecast can help us to figure it out, although it will be more work than above. Let's start with the observation equation:

$$\begin{aligned} y_t &= \ell_{t-1} + b_{t-1} + \varepsilon_t \\ y_{T+2} &= \ell_{T+1} + b_{T+1} + \varepsilon_{T+2} \\ E(y_{T+2}|Y) &= E(\ell_{T+1}|Y) + E(b_{T+1}|Y) + E(\varepsilon_{T+2}|Y) \\ E(y_{T+2}|Y) &= E(\ell_{T+1}|Y) + E(b_{T+1}|Y) \end{aligned}$$

As usual, the error term dropped out, since the error has mean 0. However, we appear to be stuck, since we need the expected value of ℓ_{T+1} and b_{T+1} we only have estimates through ℓ_T and b_T. Let's try to use the level and trend equations to find these values. First, the level

equation:

$$\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$$

$$\ell_{T+1} = \ell_T + b_T + \alpha\varepsilon_{T+1}$$

Taking the expected value of both sides:

$$E(\ell_{T+1}|Y) = E(\ell_T|Y) + E(b_T|Y) + E(\alpha\varepsilon_{T+1}|Y)$$

Since we can treat α as a constant, we can pull it outside of the expected value:

$$E(\ell_{T+1}|Y) = E(\ell_T|Y) + E(b_T|Y) + \alpha E(\varepsilon_{T+1}|Y)$$

$$E(\ell_{T+1}|Y) = E(\ell_T|Y) + E(b_T|Y) + 0$$

Note that we know both $E(\ell_T|Y)$ and $E(b_T|Y)$ — these are our in-sample estimates of ℓ_T and b_T. We have:

$$E(\ell_{T+1}|Y) = \ell_T + b_T$$

Substituting this into the observation equation from earlier, we now have:

$$E(y_{T+2}|Y) = E(\ell_{T+1}|Y) + E(b_{T+1}|Y)$$

$$E(y_{T+2}|Y) = \ell_T + b_T + E(b_{T+1}|Y)$$

Now let's use the trend equation to find $E(b_{T+1}|Y)$:

$$b_t = b_{t-1} + \alpha\beta\varepsilon_t$$

$$b_{T+1} = b_T + \alpha\beta\varepsilon_{T+1}$$

Taking the expected value of both sides (and skipping steps similar to the ones we performed

in the level equation), we have:

$$E(b_{T+1}|Y) = b_T + \alpha\beta E(\varepsilon_{T+1}|Y)$$
$$E(b_{T+1}|Y) = b_T$$

Plugging into the observation equation:

$$E(y_{T+2}|Y) = \ell_T + b_T + E(b_{T+1}|Y)$$
$$E(y_{T+2}|Y) = \ell_T + b_T + b_T$$
$$E(y_{T+2}|Y) = \ell_T + 2b_T$$
$$\hat{y}_{T+2|T} = \ell_T + 2b_T$$

It took a bit of algebra, but we found the answer. The optimal two-period ahead forecast is given by $\hat{y}_{T+2|T} = \ell_T + 2b_T$

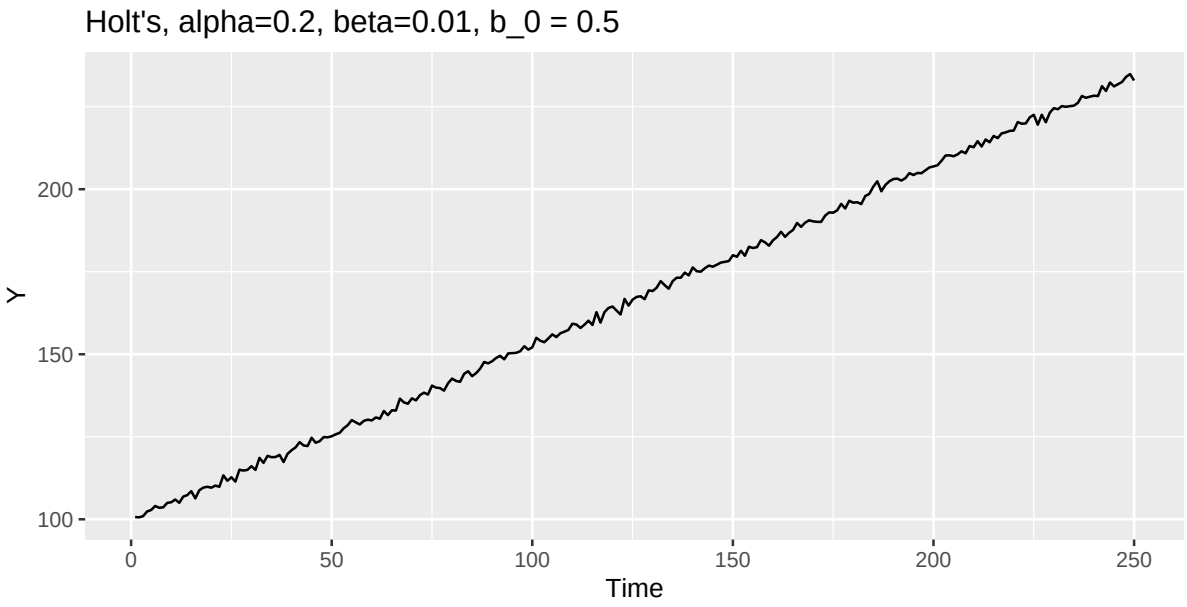
We could repeat this process for $\hat{y}_{T+3|T}$, and I encourage you to do so. If you did this, the pattern would probably become obvious. Using Holt's method under squared error loss, the optimal h-period ahead forecast is given by:

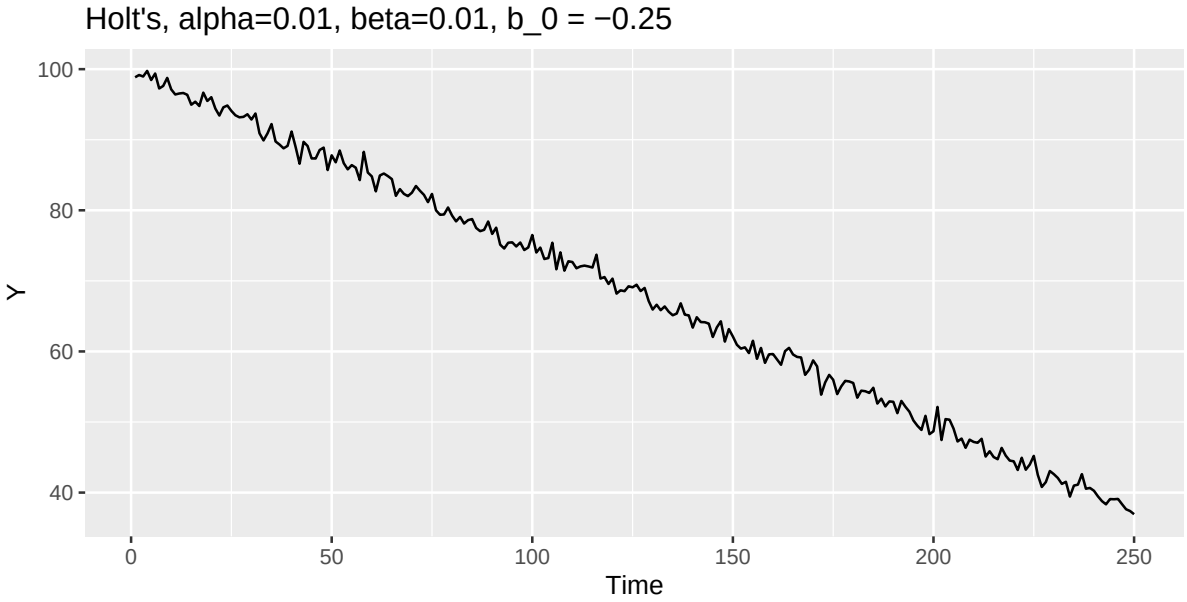
$$y_{T+h|T} = \ell_T + hb_T$$

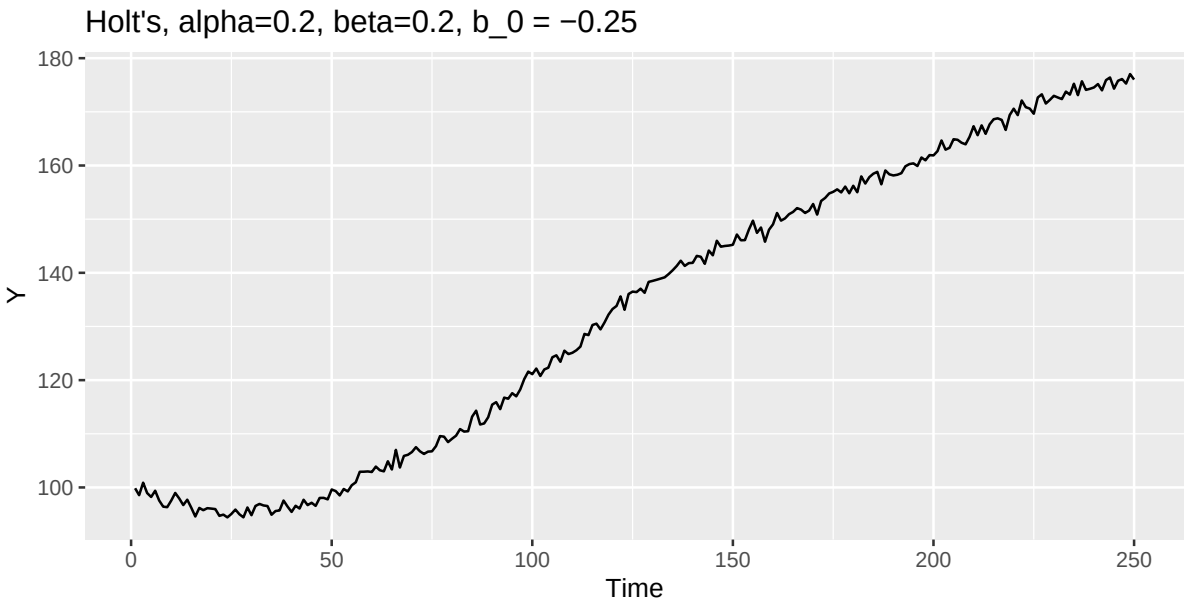
Although this takes a lot of work to show mathematically, it makes a lot of intuitive sense. Since the last estimated linear trend is b_T, we expect the series to increase by b_T units each period in the future. Therefore, the best forecast for h periods from now is this period's estimated level, plus h applications of the trend.

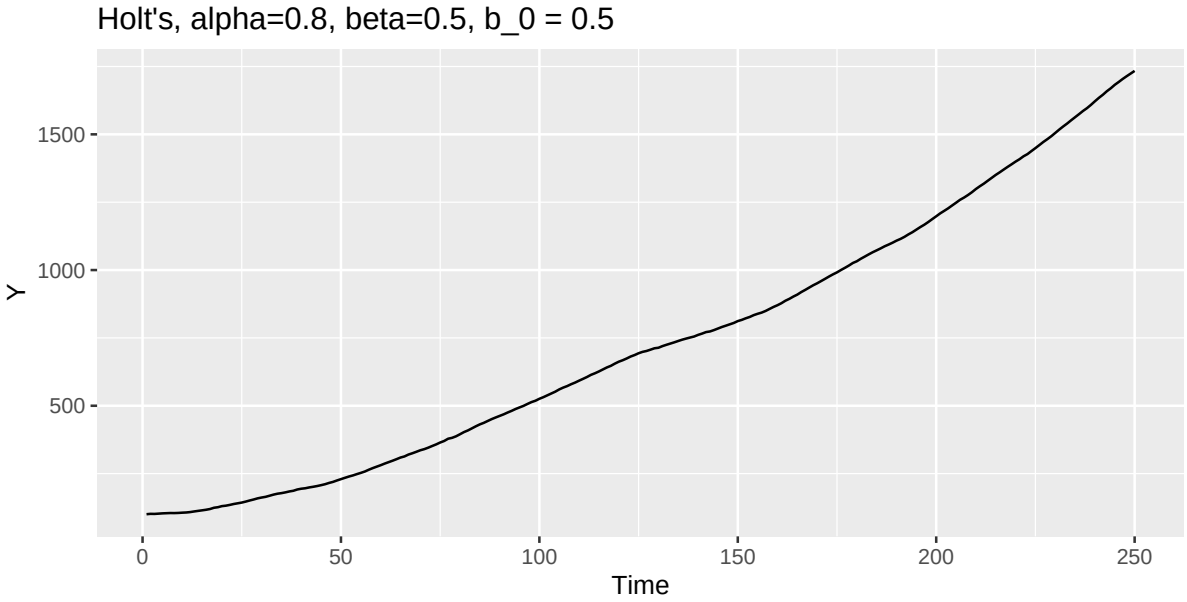
3.3 When to Use Holt's Method

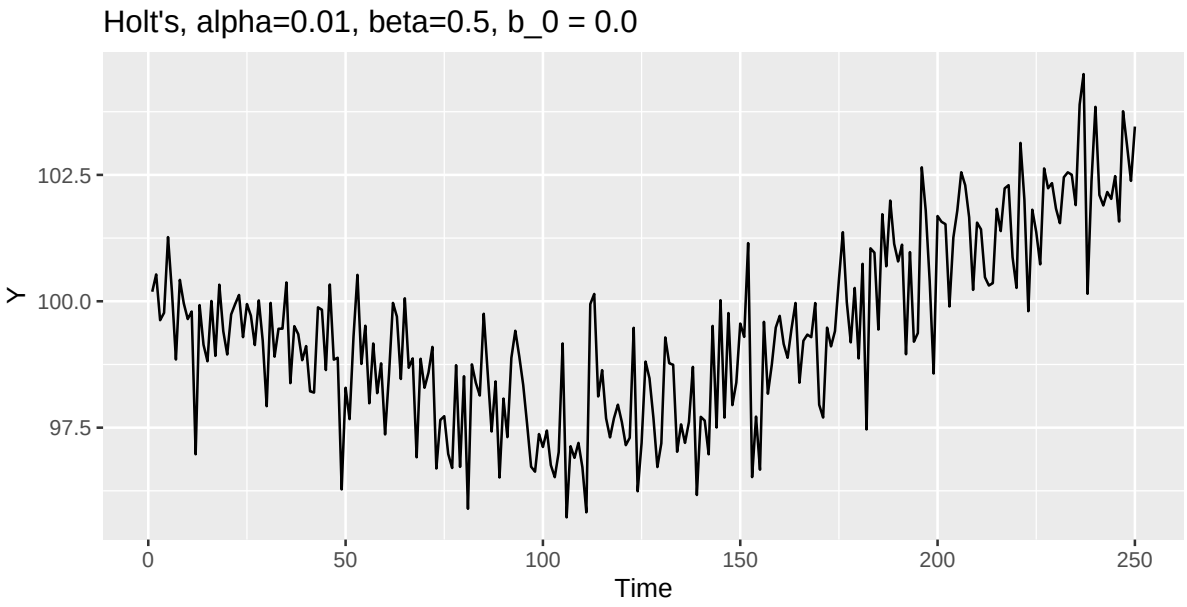
You should use Holt's method if your data has a trend, but does NOT have seasonality. Holt's method is incredibly flexible, since it can allow for a very general level process (via the level equation) and a very general trend process, with the trend allowed to change over time. Below, I have randomly generated data according to Holt's method with known parameter values. This should give you some sense of when Holt's method is appropriate.

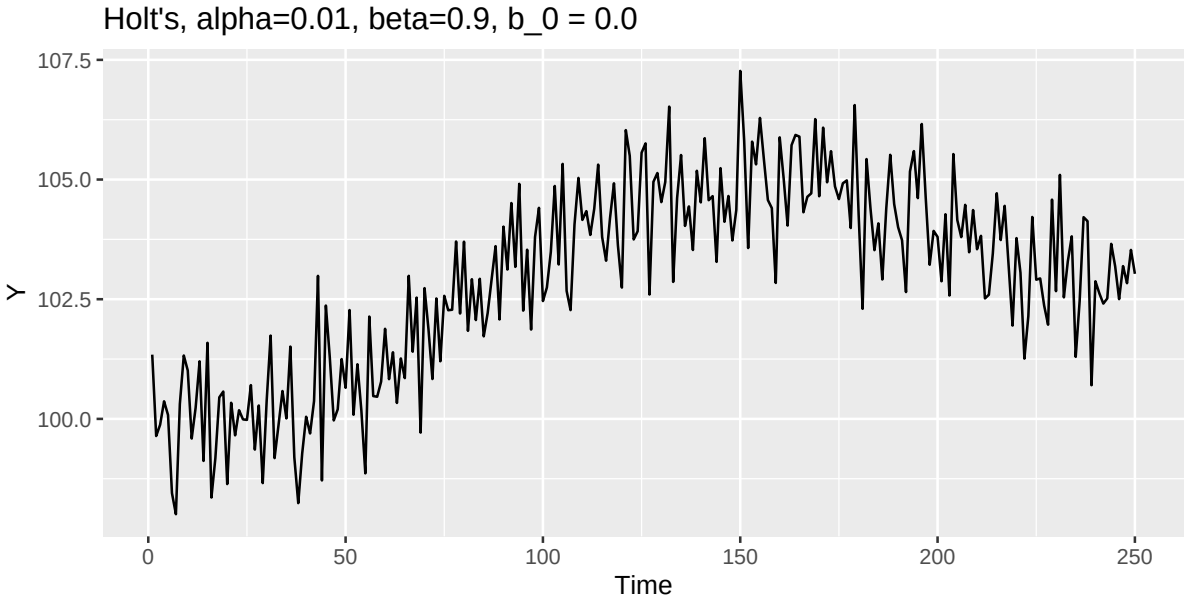

Holt's, alpha=0.2, beta=0.01, b_0 = 0.5
Y
100
150
200
0
50
100
150
200
250
Time


Holt's, alpha=0.01, beta=0.01, b_0 = −0.25
Y
40
60
80
100
0
50
100
150
200
250
Time


Holt's, alpha=0.2, beta=0.2, b_0 = −0.25
180
160
140
120
100
Y
0
50
100
150
200
250
Time


Holt's, alpha=0.8, beta=0.5, b_0 = 0.5
1500
1000
500
Y
0
50
100
150
200
250
Time





It should be apparent that by changing the parameters of Holt's model, we can capture a wide variety of time-series data — data that has much different looking trends and even trends that reverse signs. However, note that when using the Holt model to forecast, the uncertainty associated with your point forecast typically grows rapidly as the forecast horizon increases.

4 Holt-Winters' Exponential Smoother

In 1960, a student of Holt's, Peter Winters, extended Holt's method by building an exponential smoother that allows for:

1. Changes in level

2. Trend & Changes in Trend

3. Seasonality & Changes in Seasonality

Therefore, the Holt-Winters' (HW) model allows estimation and forecasting with time-series data that has seasonal patterns. This model adds an additional equation to Holt's method, so there are now four main equations (along with the error distribution). In "Component" form, we have:

$$\begin{aligned}
\text{Observation: } & y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t \\
\text{Level: } & \ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \\
\text{Trend: } & b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \\
\text{Seasonal: } & s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \\
\text{Error: } & \varepsilon_t \sim N(0, \sigma)
\end{aligned}$$

where m is the number of seasons in a year (e.g. $m = 12$ if the data is monthly) and γ is the seasonal smoothing parameter.

From the observation equation we can rearrange to solve for the error term at time t:

$$\varepsilon_t = y_t - \ell_{t-1} - b_{t-1} - s_{t-m}$$

We can use this to find the "Error Correction" form. I will not go through the math, but it is written as:

$$\begin{aligned}
\text{Observation: } & y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t \\
\text{Level: } & \ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t \\
\text{Trend: } & b_t = b_{t-1} + \alpha\beta\varepsilon_t \\
\text{Seasonal: } & s_t = s_{t-m} + \gamma\varepsilon_t \\
\text{Error: } & \varepsilon_t \sim N(0, \sigma)
\end{aligned}$$

While this series of equations looks substantially more complicated then the SES equations, the intuition is basically the same. Every time a new data point is recorded, we update our

estimate of the level, the trend, and the seasonal component for the season in which the new data point was recorded. The level, trend, and each of the seasonal components can all evolve over time.

Like the previous two methods, we typically need to estimate the smoothing parameters: α, β, and γ, as well as the initial level, ℓ_0 and initial trend, b_0. Additionally, we need to estimate initial conditions for all of the seasons: $s_{-m+1}, s_{-m+2}, \cdots, s_0$. The increase in the number of parameters compared to Holt's method can be substantial. For example, if our data is observed daily, this means we would need to estimate 365 additional parameters. Therefore, when using this seasonal method, it is best if you have a relatively long data series (at least 10-20 years of data regardless of the seasonal frequency), so that you have enough data points to reliably estimate all of the initial conditions. Once we have estimated all the parameters, we can use the equations of the model to form estimates of $\ell_1, \ell_2, \cdots, \ell_T$, $b_1, b_2, \cdots, b_T$, and $s_1, s_2, \cdots, s_T$.

5 ETS Notation

After inspecting a data series graphically, it may seem fairly obvious which method is most appropriate to use. For example, if it looks like there is a strong trend and no seasonality, we should probably use Holt's method. However, there are cases we didn't consider. For example, what if it looks like there is seasonality, but no trend? Or what if it looks like there may be a slight trend over time, so you can't tell if including a trend is appropriate?

One possible solution to this problem is to fit all possible exponential smoothing models, and choose the one that has the lowest AICc. In R, there is a built in function that does this very quickly, so it is just as easy as using any of the methods individually. In order to understand what R is doing, and in order to classify all of the possible exponential smoothers, we will introduce "**ETS Notation**".

ETS stands for Error, Trend, Seasonal. All of these three components can take on at least two values: "A" for Additive or "M" for multiplicative. For example, if the error term is Additive, this means that the error term is best expressed as the number of units the data is measured in. If the error term is Multiplicative, this means that the error term is best expressed as a percentage. For a more concrete example, consider the case of a time-series, Y, that has a mean (level) of 100. Suppose the error at time period t was 5. If the error term was Additive, we would say this was an error of 5. If it was multiplicative, we would say this was an error of

$5/100 = 0.05 = 5\%$.

All exponential smoothers will have an error term that will be either additive or multiplicative. However, not all smoothers will have a trend or seasonality. Therefore, trend and seasonality can take values of "N" for None. For example, the SES model we introduced has additive errors and no trend or seasonality so we would write it as (E,T,S)=(A,N,N). Holt's method had additive errors and an additive trend, we could write it as (E,T,S)=(A,A,N).

The trend component can also be "damped", meaning that while there is a trend in-sample, we expect it to die out slowly over the forecast horizon. Therefore, the trend component can take on values of: "N", "A", "M", "Ad", or "Md", where "Ad" stands for Additive damped and "Md" stands for Multiplicative damped.

5.1 Examples

Model	(E,T,S)
SES	(A,N,N)
Holt	(A,A,N)
Holt-Winters'	(A,A,A)

A model that has no trend, so it is relatively flat over time, but does have seasonality, could be written as: (A,N,A) if the error & seasonality was additive, or (M,N,M) if the error & seasonality was multiplicative.

A model with additive errors, an additive damped trend, but no seasonality would be written as (A,Ad,N).

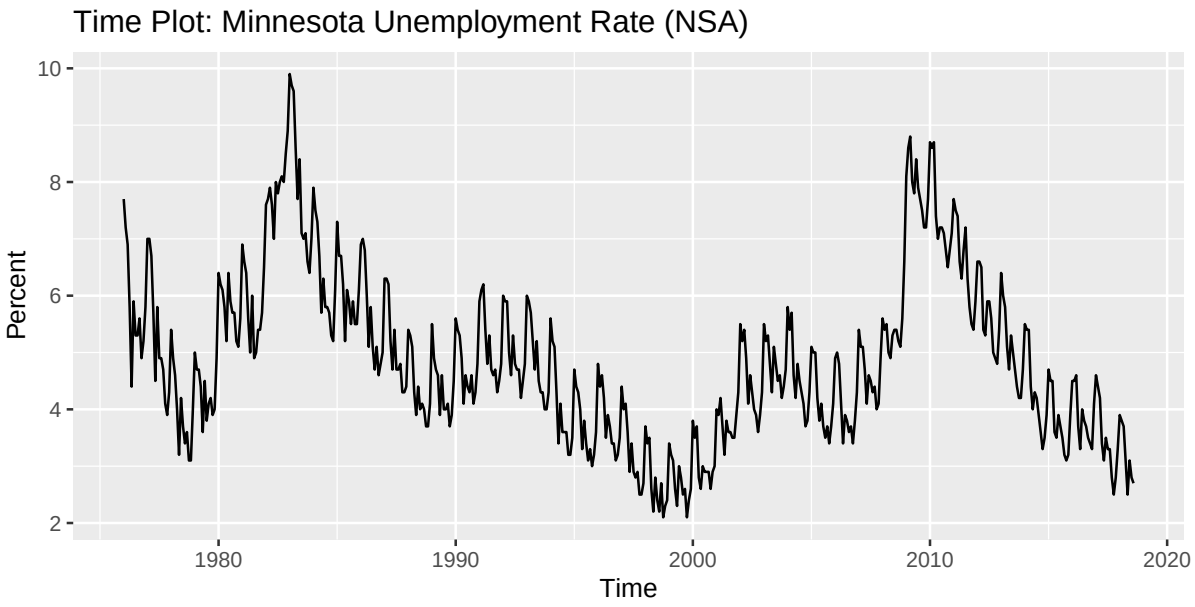
A model with additive errors, an additive damped trend, and additive seasonality would be written as (A,Ad,A).

Clearly, there are many more possibilities. Again, there is a function in R that, given a data set, will run through all possibilities and return the one that has the smallest AICc (and so is likely, although not guaranteed, to forecast the best).

6 Application: MN Unemployment Rate

Let's consider the non-seasonally adjusted Minnesota Unemployment Rate. We will use data through August 2018, and then forecast 12 periods ahead. At the end, we will compare the forecasts to what has actually happened between September 2018 and August 2019.

```r
# Load Data
data_set <- fredr(series_id = "MNURN")
Y <- data_set$value
Y <- ts(Y,start=c(1976,1),frequency=12)
Y <- window(Y,end=c(2018,8))
# Time Plot
autoplot(Y) +
  ggtitle("Time Plot: Minnesota Unemployment Rate (NSA)") +
  ylab("Percent")
```



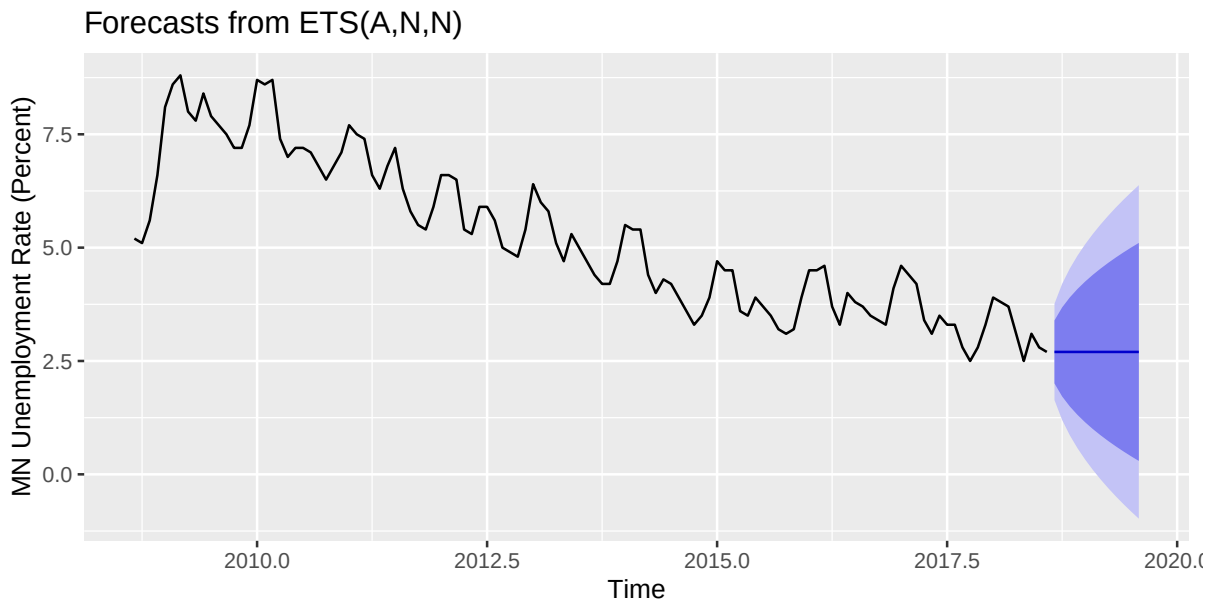
Let's form $H = 12$ period-ahead forecasts using each of the three methods.

6.1 Simple Exponential Smoothing

In R, the commands to forecast with exponential smoothing are fairly straightforward. Let's start with SES:

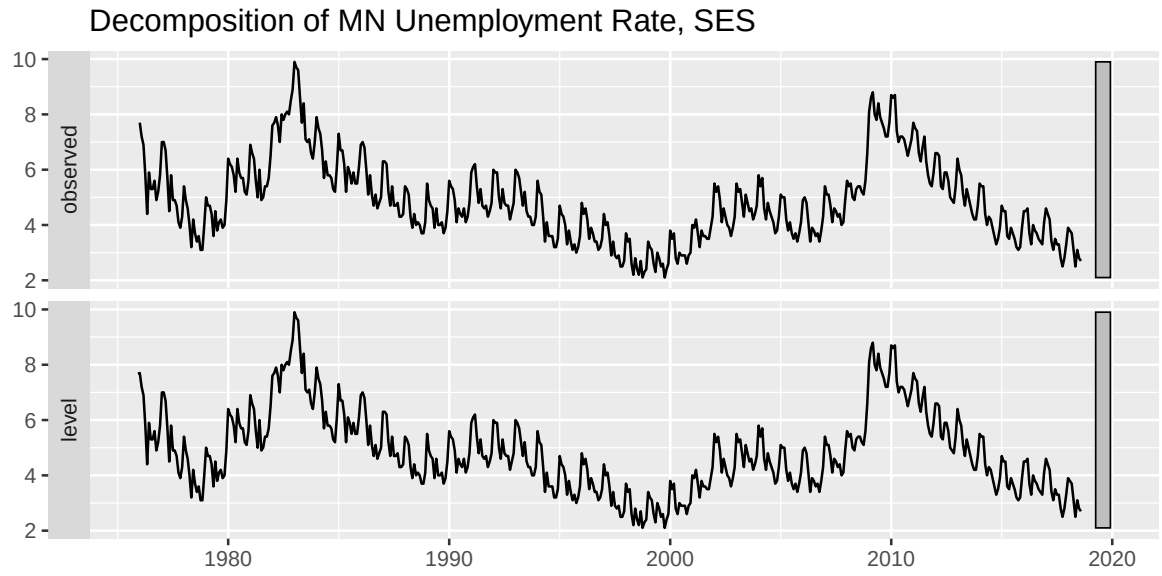
```r
# Fit SES Model
fit_ses <- ets(Y,model="ANN") #Additive error, No trend, No seasonality
# Forecast from SES Model
fcst_ses <- forecast(fit_ses,h=12)
# Time plot of data & forecasts
```

```
autoplot(fcst_ses,include=120) + ylab("MN Unemployment Rate (Percent)")
```

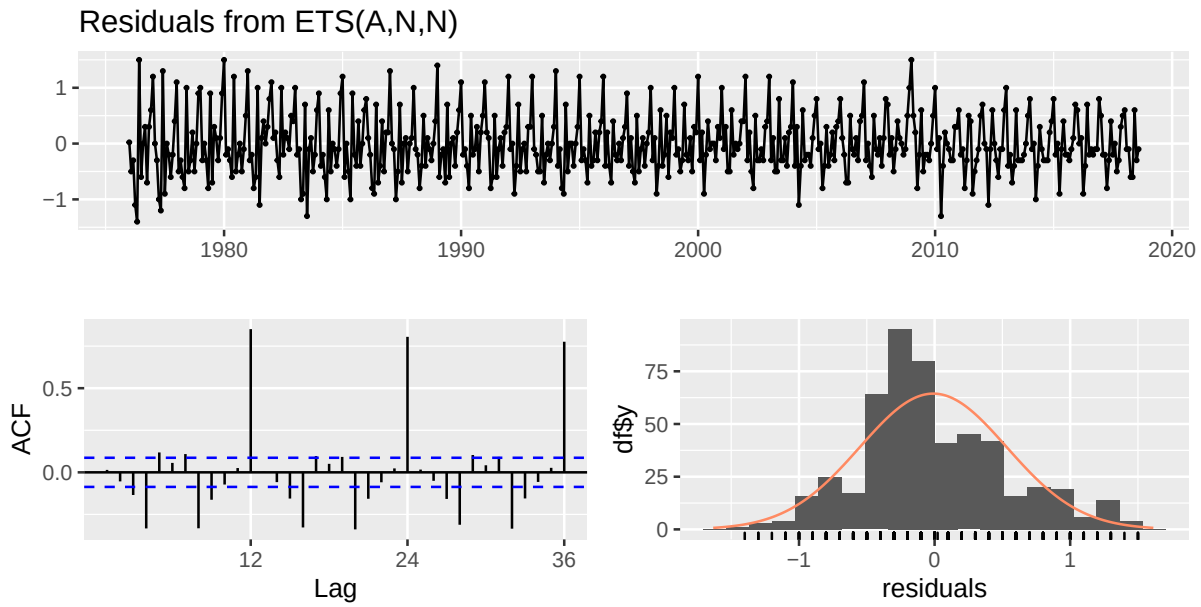


Recall that as a byproduct of SES, we have the estimated level at each point in time in-sample. We can view the actual data and estimated level by autoplotting the fit. We may also wish to inspect the residuals. Note that there are clear seasonal patterns in the residuals, as the SES model is not able to capture seasonality.

```
# Plot
autoplot(fit_ses) +
  ggtitle("Decomposition of MN Unemployment Rate, SES")
```



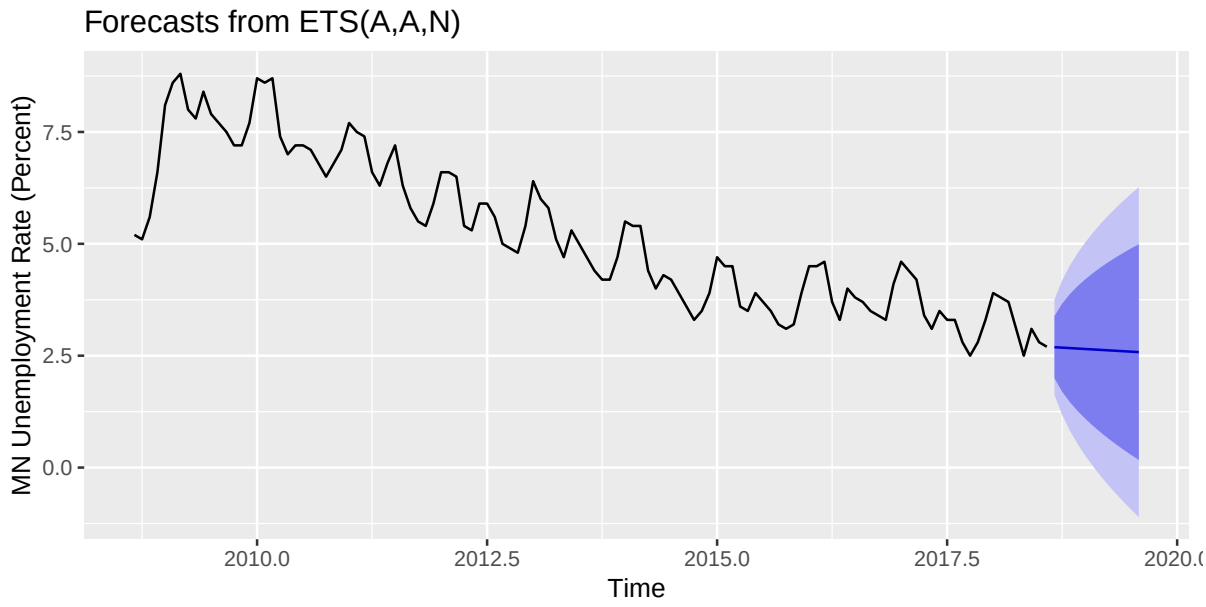
```r
# Check residuals

checkresiduals(fit_ses)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(A,N,N)
## Q* = 1048.3, df = 22, p-value < 2.2e-16
##
## Model df: 2.   Total lags used: 24
```

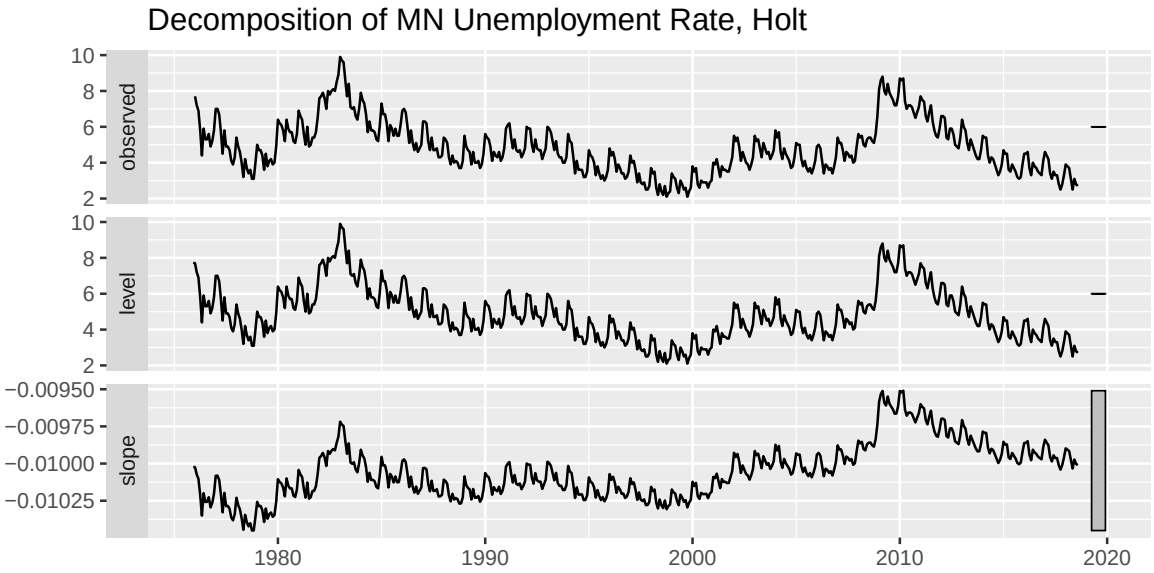
6.2 Holt's Method

```
# Fit with Holt's method
fit_holt <- ets(Y,model="AAN",damped=FALSE) # Additive level and trend, No seasonal
# Forecast with Holt's method
fcst_holt <-forecast(fit_holt,h=12)
# Time plot of data & forecasts
autoplot(fcst_holt,include=120) + ylab("MN Unemployment Rate (Percent)")
```

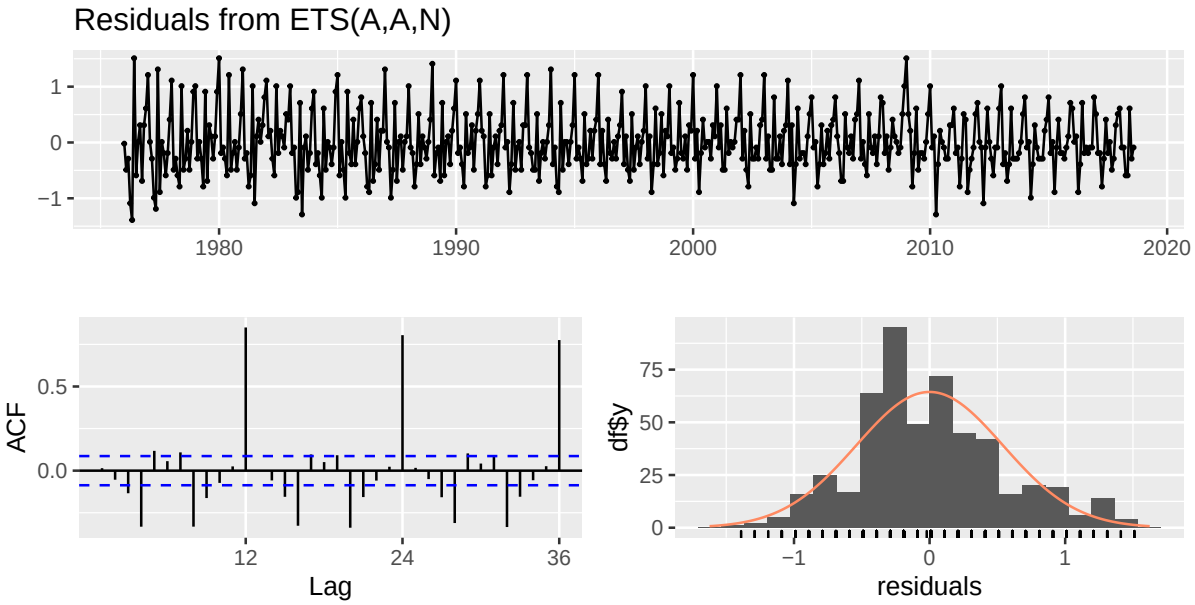


We now have an estimated level and trend. We can decompose our data into three parts: level, trend, and error. Note that there are clear seasonal patterns in the residuals.

```
# Plot
  autoplot(fit_holt) +
  ggtitle("Decomposition of MN Unemployment Rate, Holt")
```



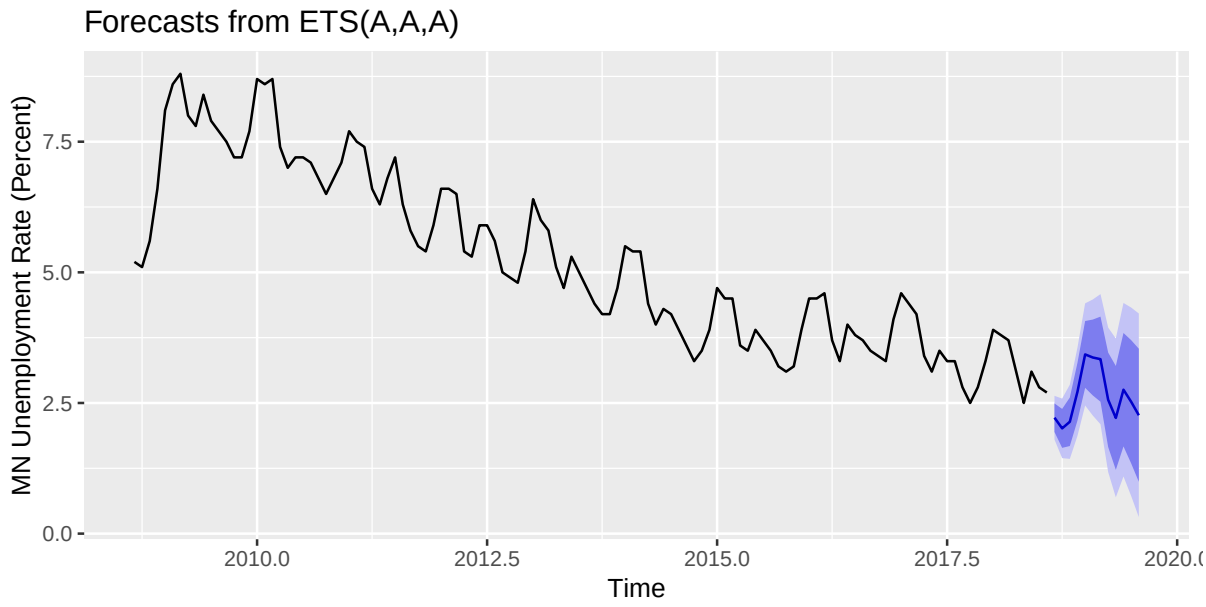

```
# Check Residuals

checkresiduals(fit_holt)
```



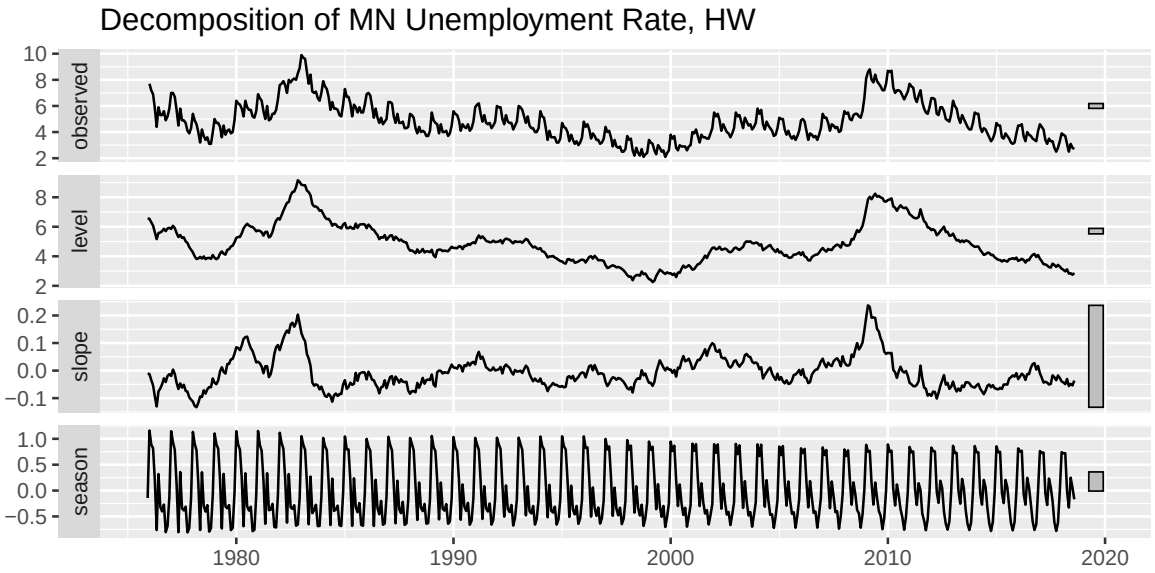

```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(A,A,N)
## Q* = 1047, df = 20, p-value < 2.2e-16
##
## Model df: 4.   Total lags used: 24
```

6.3 Holt-Winters' Method

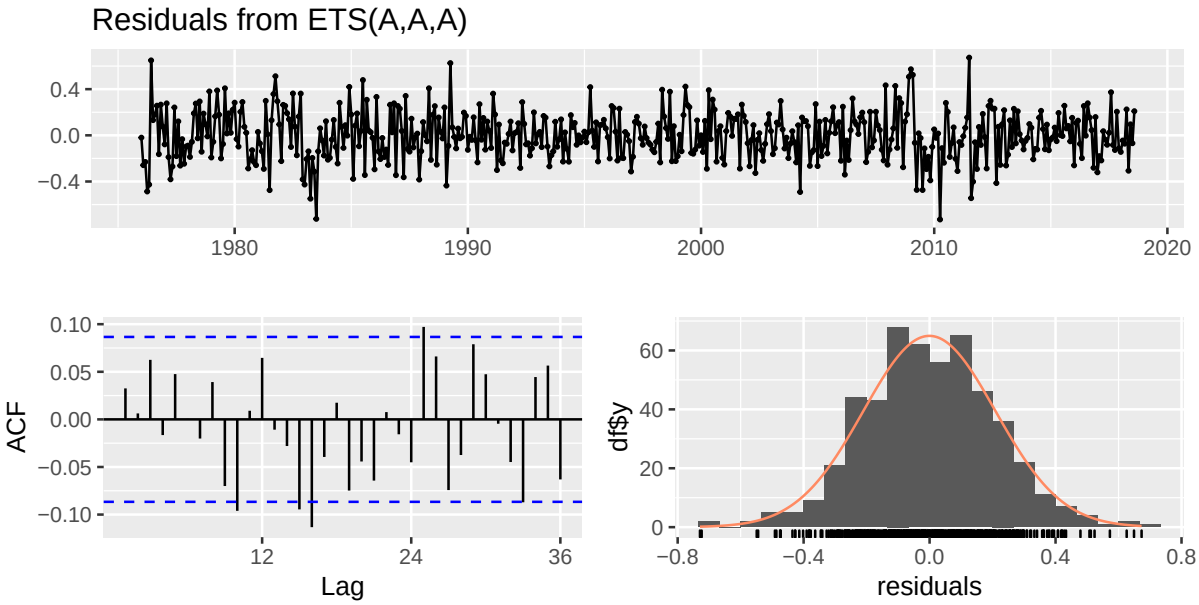
```
# Fit with Holt-Winters'
fit_hw <- ets(Y,model="AAA",damped=FALSE)
# Forecast with Holt-Winters'
fcst_hw <- forecast(fit_hw,h=12)
# Time plot of data & forecasts
autoplot(fcst_hw,include=120) + ylab("MN Unemployment Rate (Percent)")
```




```
autoplot(fit_hw) +
  ggtitle("Decomposition of MN Unemployment Rate, HW")
```




```
# Check Residuals

checkresiduals(fit_hw)
```

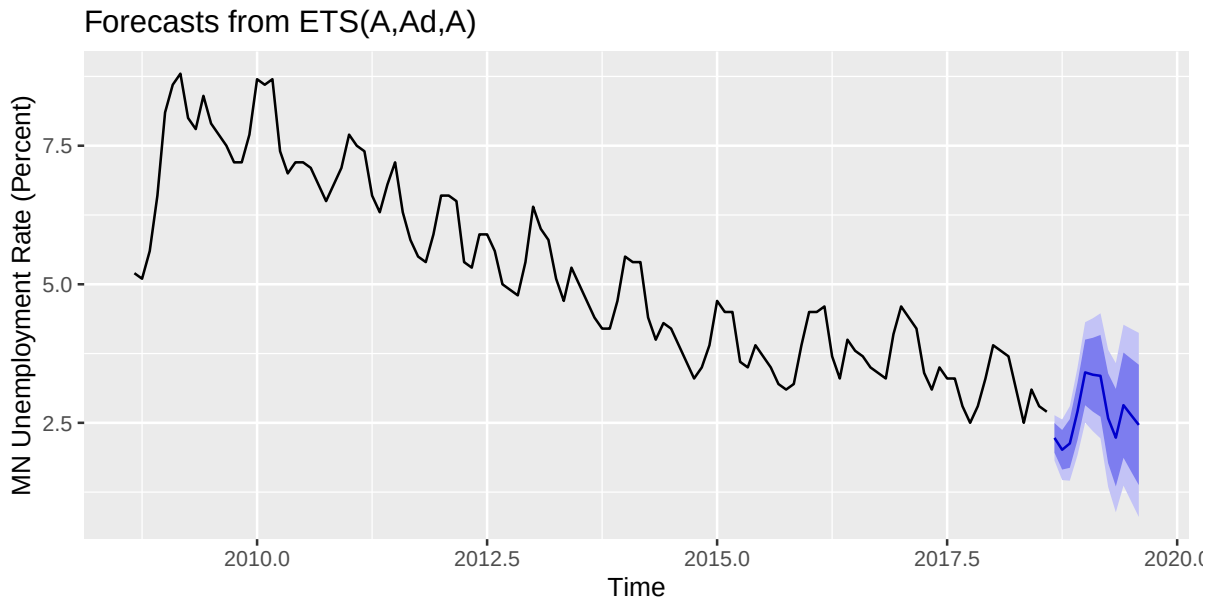



```
##
##  Ljung-Box test
##
## data:  Residuals from ETS(A,A,A)
## Q* = 35.074, df = 8, p-value = 2.592e-05
##
## Model df: 16.   Total lags used: 24
```

Note that the residuals no longer have a seasonal pattern, as the HW method is able to account for the seasonality, and includes adjustments for seasonality it in it's in-sample predictions.

6.4 Automatic ETS

```
# Have R select the ETS model with the smallest AICc
fit <- ets(Y)
# Use that model to forecast
fcst <- forecast(fit,h=12)
# Time plot of data & forecasts
autoplot(fcst,include=120) + ylab("MN Unemployment Rate (Percent)")
```



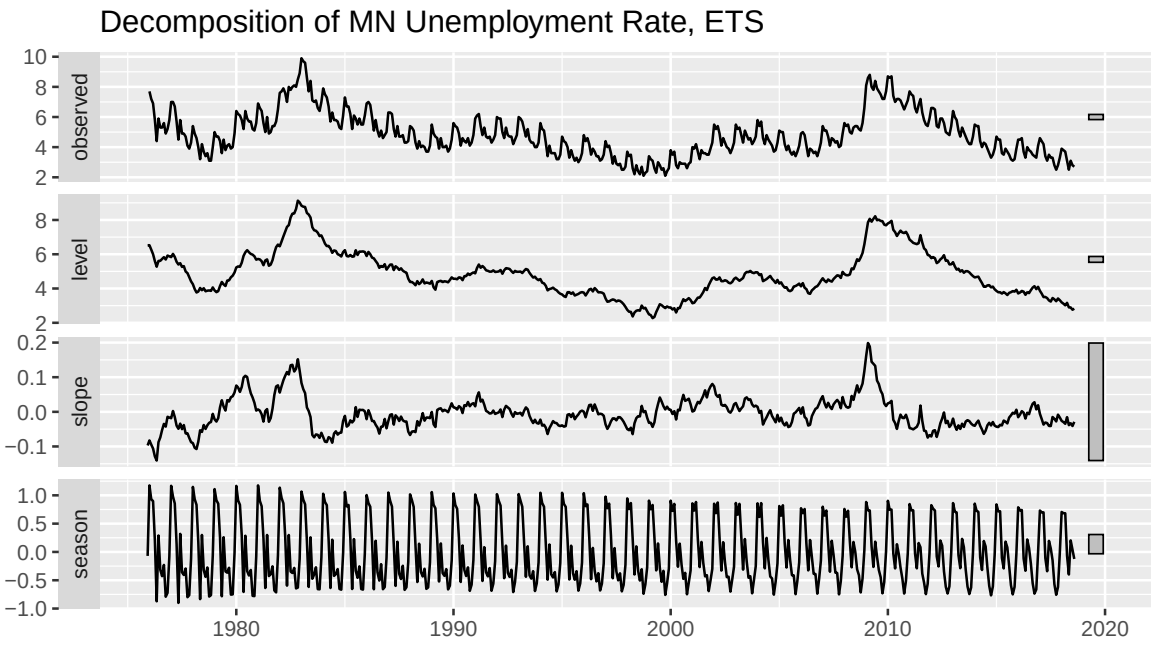
And check to see which exponential smoother was chosen by the computer (i.e., which one had the smallest AICc):

```
fit$method
```

```
## [1] "ETS(A,Ad,A)"
```

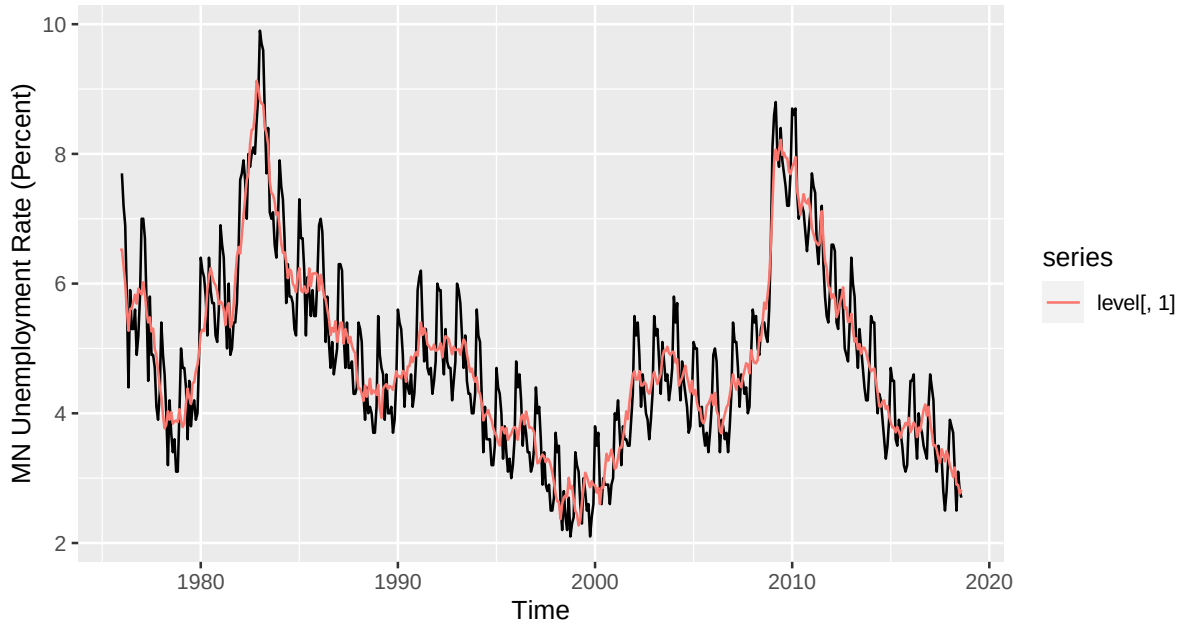
So we can see that the method that fits the data the best is a model with additive error, an additive damped trend, and additive seasonality.

```
autoplot(fit) +
  ggtitle("Decomposition of MN Unemployment Rate, ETS")
```



We can also plot the estimated level on the same time plot as the data, to see how they compare.

```
# Get the estimated level AND trend
level <- fcst$model$states[,1:3]
autoplot(Y) + autolayer(level[,1]) + ylab("MN Unemployment Rate (Percent)")
```

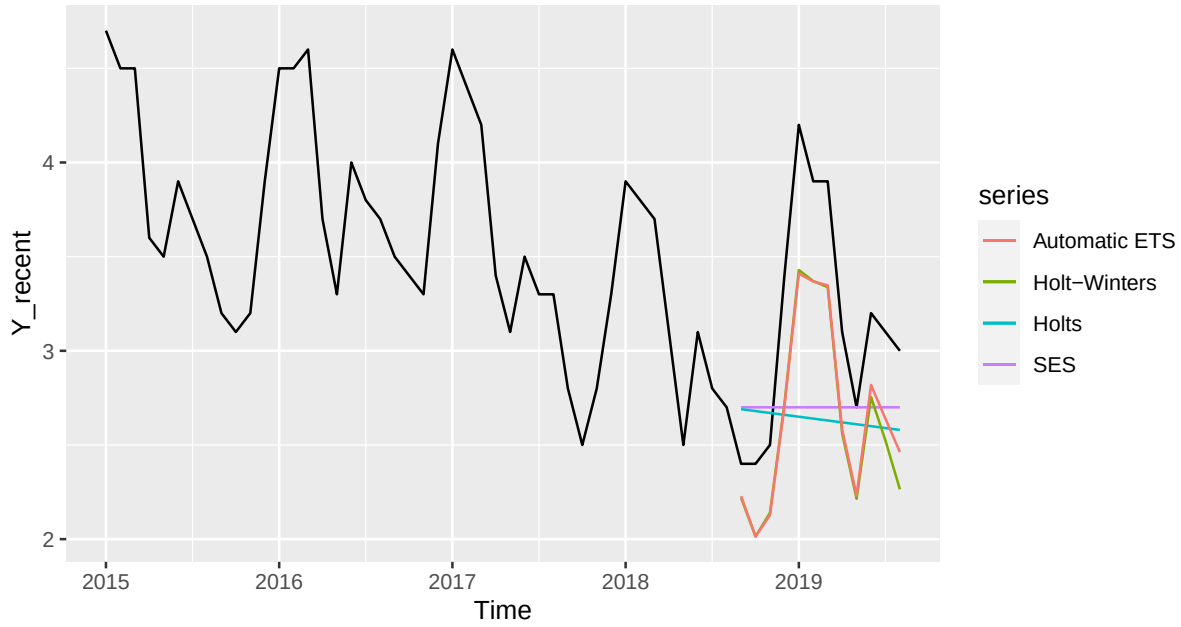


Note that the level acts like a seasonally adjusted version of the data, since the seasonality is stripped away.

6.5 Forecast Performance

We have made four forecasts: SES, Holt's, HW, and automatic ETS (Holt's with damped trend). Let's see how they compared to what actually happened.

```
data_set <- fredr(series_id = "MNURN")
Y_recent <- ts(data_set$value,start=c(1976,1),frequency = 12)
Y_recent <- window(Y_recent,start=c(2015,1),end=c(2019,8))


autoplot(Y_recent) +
  autolayer(fcst_ses$mean, series="SES") +
  autolayer(fcst_holt$mean, series="Holts") +
  autolayer(fcst_hw$mean, series="Holt-Winters") +
  autolayer(fcst$mean, series="Automatic ETS")
```



Since we know what actually happened, we can calculate the RMSFE from each method. We could do this by hand, or use the "accuracy" function in RStudio by giving the function both the forecasts and the true values. The RMSFE from each model is presented below:

Model	RMSFE
SES	0.74
Holt's	0.78
Holt Winters'	0.55
Auto ETS	0.51

From the graph, we ca see that all methods tended to under-forecast the unemployment rate, expecting it to continue falling. Of the four methods, Holt-Winters' and the automatic ETS methods performed best according to RMSFE, each missing the actual unemployment rate by roughly 0.5 percentage points each period. In large part, the relative accuracy of HW and the automatic ETS methods is driven by their ability to capture the strong seasonality in the series. Therefore, both correctly forecast that the unemployment rate would be relatively low late in 2018, relatively high in January and February of 2019, etc.