Stat 443. Time Series and Forecasting.

Key ideas:
From forecasting rules to time series models assuming additive innovations,
forecast is the conditional expectation of future observation given the observed past;
prediction interval assuming normally distributed innovations;
exponential smoothing.

Important: interpret equations so that you can easily go back and forth between the verbal explanations and mathematical expressions.

Notation: $h$ is a positive integer.
$\widehat{y}_{t+1|t}$ is the 1-step forecast for $Y_{t+1}$ given observed $y_1, \ldots, y_t$.
$\widehat{y}_{t+h|t}$ is the $h$-step forecast for $Y_{t+h}$ given observed $y_1, \ldots, y_t$.

Consider stochastic models for times series data with **additive** innovation (or disturbance or noise). Why called innovation?
In this case, if the forecast of $Y_{t+1}$ given $y_1, \ldots, y_t$ is a function $\widehat{g}_t(y_1, \ldots, y_t)$, then an additive stochastic model implies that there is an innovation rv $\epsilon_{t+1}$ independent of the past such that

$$Y_{t+1} = g_t(Y_1, \ldots, Y_t) + \epsilon_{t+1}, \quad \mathsf{E}\,[Y_{t+1}] = \mathsf{E}\,[g_t(Y_1, \ldots, Y_t)],$$

where $\mathsf{E}\,(\epsilon_{t+1}) = 0$. Then a 1-step forecast is the conditional expectation of future observation given the observed past:

$$\begin{aligned}
&\mathsf{E}\,[Y_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] \\
&\quad = \mathsf{E}\,[g_t(Y_1, \ldots, Y_t) + \epsilon_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] \\
&\quad = \mathsf{E}\,[g_t(y_1, \ldots, y_t) + \epsilon_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] \\
&\quad = g_t(y_1, \ldots, y_t) + \mathsf{E}\,[\epsilon_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] \\
&\quad = g_t(y_1, \ldots, y_t) + \mathsf{E}\,[\epsilon_{t+1}] = g_t(y_1, \ldots, y_t) + 0 \\
\widehat{y}_{t+1|t} &\quad = \widehat{\mathsf{E}}\,[Y_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] = \widehat{g}_t(y_1, \ldots, y_t)
\end{aligned}$$

$\widehat{\mathsf{E}}$ and $\widehat{g}$ mean that there may be estimated parameters in $g_t$ (such as intercept and slope for linear in previous observation).
Similarly $\widehat{y}_{t+2|t}$ is based on $\mathsf{E}\,[Y_{t+2}|Y_1 = y_1, \ldots, Y_t = y_t]$.

## Cases of average in training set, persistence, and linear in most recent observation

In general, there is a parameter $\theta$ for $g_t$, so write forecast rule as $g_t(Y_1,\ldots,Y_t,\theta)$, where $\theta$ is estimated from the training set $y_1,\ldots,y_n$ to get $\widehat{\theta}$, and might be updated as more observations are obtained.

For $t > n$, the 1-step forecast is

$$\widehat{y}_{t+1|t} = \widehat{g}_t(y_1,\ldots,y_t) = g_t(y_1,\ldots,y_t;\widehat{\theta}).$$

The stochastic model is taken as

$$Y_{t+1} = g_t(Y_1,\ldots,Y_t;\theta) + \epsilon_{t+1}, \qquad (*)$$

where $\epsilon_{t+1}$ has mean 0 and is independent of the past, and $\{\epsilon_i\}$ is an iid (independent and identically distributed) sequence.

$$\widehat{y}_{t+1|t} = \widehat{\mathsf{E}}\left[Y_{t+1}|Y_1 = y_1,\ldots,Y_t = y_t\right] = g_t(y_1,\ldots,y_t;\widehat{\theta})$$

Forecast rules (see Section 3.1 of H&A): $g_t(Y_1, \ldots, Y_t, \theta)$ with $\theta$ to be estimated from training set.

1. Average of past observations in training set of size $n$, assuming iid (independent and identically distributed):

$$\widehat{y}_{t+1|t} = g_t(y_1, \ldots, y_t; \widehat{\mu}) = n^{-1}(y_1 + \ldots + y_n) = \widehat{\mu}, \quad t > n,$$

$Y_{t+1} = g_t(Y_1, \ldots, Y_t; \mu) + \epsilon_{t+1} = \mu + \epsilon_{t+1}, \quad t > n,$ stochastic using (*)

$\theta = \mu$ and $\mathrm{E}\left[g_t(Y_1, \ldots, Y_n, \theta)\right] = \mu = \mathrm{E}\left[Y_i\right]$, where $\epsilon_{t+1}$ has mean 0 and is independent of the past. That is, the stochastic model is

$$Y_i = \mu + \epsilon_i, \quad i = 1, \ldots,$$

for an iid sequence of $\{Y_i\}$ (or iid sequence $\{\epsilon_i\}$ with mean 0). This model is called "white noise".

Exercise: What is the standard error that can be used for prediction intervals?

2. Persistence $g_t(Y_1, \ldots, Y_t) = Y_t$ (no $\theta$) and, with additive innovation, from equation (*),

$$\widehat{y}_{t+1|t} = y_t, \quad Y_{t+1} = Y_t + \epsilon_{t+1}, \quad t \geq 1,$$

where $\epsilon_{t+1}$ has mean 0 and is independent of $Y_1, \ldots, Y_t$. This model is called a random walk because the next observation is the previous one plus some random variable with mean 0.

Exercise: What is the standard error that can be used for prediction intervals?

3. Autoregressive $g_t(Y_1, \ldots, Y_t; \theta)$ where $\theta = (\mu, \phi_1, \ldots, )$.
$Y_t$ is the sum of a linear function of $Y_{t-1}, \ldots, Y_{t-p}$ with "noise", where $p$ is a positive integer. With $p = 1$, and $t \geq 1$, from equation (*), $g_t(y_1, \ldots, y_t; \widehat{\mu}, \widehat{\phi}_1) = \widehat{\mu} + \widehat{\phi}_1(y_t - \widehat{\mu})$,

$$\widehat{y}_{t+1|t} = \widehat{\mu} + \widehat{\phi}_1(y_t - \widehat{\mu}), \quad Y_{t+1} = \mu + \phi_1(Y_t - \mu) + \epsilon_{t+1}, \tag{1}$$

where $\epsilon_1, \epsilon_2, \ldots$ are iid with mean 0 and var. $\sigma_\epsilon^2$, and $\epsilon_{t+1}$ is an innovation rv indep. of $Y_t, Y_{t-1}, \ldots, Y_1$. Note that (1) is a Markov process (Markov chain of order 1) with continuous state space (if $Y$'s are continuous rv's).

$$\mathsf{E}\left[Y_{t+1} | Y_1 = y_1, \ldots, Y_t = y_t\right] = \mathsf{E}\left[Y_{t+1} | Y_t = y_t\right]$$

$$= \mu + \phi_1(y_t - \mu) + \mathsf{E}\left(\epsilon_{t+1}\right) = \mu + \phi_1(y_t - \mu).$$

The parameters $\mu, \phi_1, \sigma_\epsilon^2$ are estimated based on the training set. There is a constraint on $\phi_1$ in order than it is estimable (later slide).

For the estimable case, what is a standard error for $\widehat{y}_{t+1|t}$ for $t > n$ and that can be used for prediction intervals?

$$\begin{aligned} \text{Var}\,[Y_{t+1}|Y_1 = y_1, \ldots, Y_t = y_t] &= \text{Var}\,[Y_{t+1}|Y_t = y_t] \\ &=^{\text{why?}} \text{Var}\,(\epsilon_{t+1}) = \sigma_\epsilon^2 \end{aligned}$$

Gaussian/normality assumption that is common for further derivations, such as for prediction intervals. If $\{\epsilon_t\}$ is an sequence of iid $N(0, \sigma_\epsilon^2)$ rv's, then $Y_{t+1} = \mu + \phi_1(Y_t - \mu) + \epsilon_{t+1}$ implies

$$[Y_{t+1}|Y_t = y_t] \sim N(\mu + \phi_1(y_t - \mu), \sigma_\epsilon^2)$$

With known parameters, the 90% prediction interval is

$$\mu + \phi_1(y_t - \mu) \pm z_{0.95}\sigma_\epsilon$$

$100(1 - \alpha)$% prediction interval (for $0 < \alpha \leq 0.5$) is

$$\mu + \phi_1(y_t - \mu) \pm z_{1-\alpha/2}\sigma_\epsilon$$

| $\alpha$ | $1-\alpha$ | $z_{1-\alpha/2}$ |
|---|---|---|
| 0.5 | 0.5 | 0.675 |
| 0.4 | 0.6 | 0.842 |
| 0.3 | 0.7 | 1.036 |
| 0.2 | 0.8 | 1.282 |
| 0.1 | 0.9 | 1.645 |
| 0.05 | 0.95 | 1.960 |

Estimated 90% prediction interval is

$$\widehat{\mu} + \widehat{\phi}_1(y_t - \widehat{\mu}) \pm z_{0.95}\widehat{\sigma}_\epsilon = (\widehat{F}_{Y_{t+1}|\mathcal{F}_t}(0.05), \widehat{F}_{Y_{t+1}|\mathcal{F}_t}(0.95)$$

where $\widehat{\mu}, \widehat{\phi}_1, \widehat{\sigma}_\epsilon$ are obtained based on the training set and $\widehat{F}$ is the estimated cdf of $Y_{t+1}$ given the past $\mathcal{F}_t$.

AR(1): The cdf of $Y_{t+1}$ given the past $\mathcal{F}_t$ from above is based on $N(\mu+\phi_1(y_t-\mu), \sigma_\epsilon^2)$. In R notation, $\texttt{pnorm}(., \mu+\phi_1(y_t-\mu), \sigma_\epsilon)$. Similar steps can be applied for more complex models to come.

Special cases: exercise: verify the results below, review probability rules for linear combinations of random variables

(a) $-1 < \phi_1 = \phi < 1$: This is a condition for the Markov process (1) to have a stationary distribution. Stationarity implies that $F_{Y_i,\ldots,Y_{i+h}} = F_{Y_j,\ldots,Y_{j+h}}$ for all integers $i < j$ and $h > 0$ (distribution is invariant to shift of time index). This implies that the mean and variance of $Y_t$ do not depend on $t$. Taking means and variances of (1), one gets

$$\mathrm{E}\,(Y_{t+1}) = \mu + \phi \mathrm{E}\,(Y_t) - \phi\mu, \quad \mathrm{Var}\,(Y_{t+1}) = \phi^2 \mathrm{Var}\,(Y_t) + \sigma_\epsilon^2.$$

For weak stationarity (mean and variance stationarity) then $\mu_Y = \mathrm{E}\,(Y_{t+1}) = \mathrm{E}\,(Y_t)$, $\sigma_Y^2 = \mathrm{Var}\,(Y_{t+1}) = \mathrm{Var}\,(Y_t)$. Then one must have $\mu_Y = \mu + \phi\mu_Y - \phi\mu$ or $\mu_Y = \mu$, and $\sigma_Y^2 = \phi^2\sigma_Y^2 + \sigma_\epsilon^2$ or $\sigma_\epsilon^2 = (1-\phi^2)\sigma_Y^2$ and $-1 < \phi < 1$; (1) can be written as

$$Y_{t+1} - \mu = \phi(Y_t - \mu) + \epsilon_{t+1}$$

and then recursively (exercise)

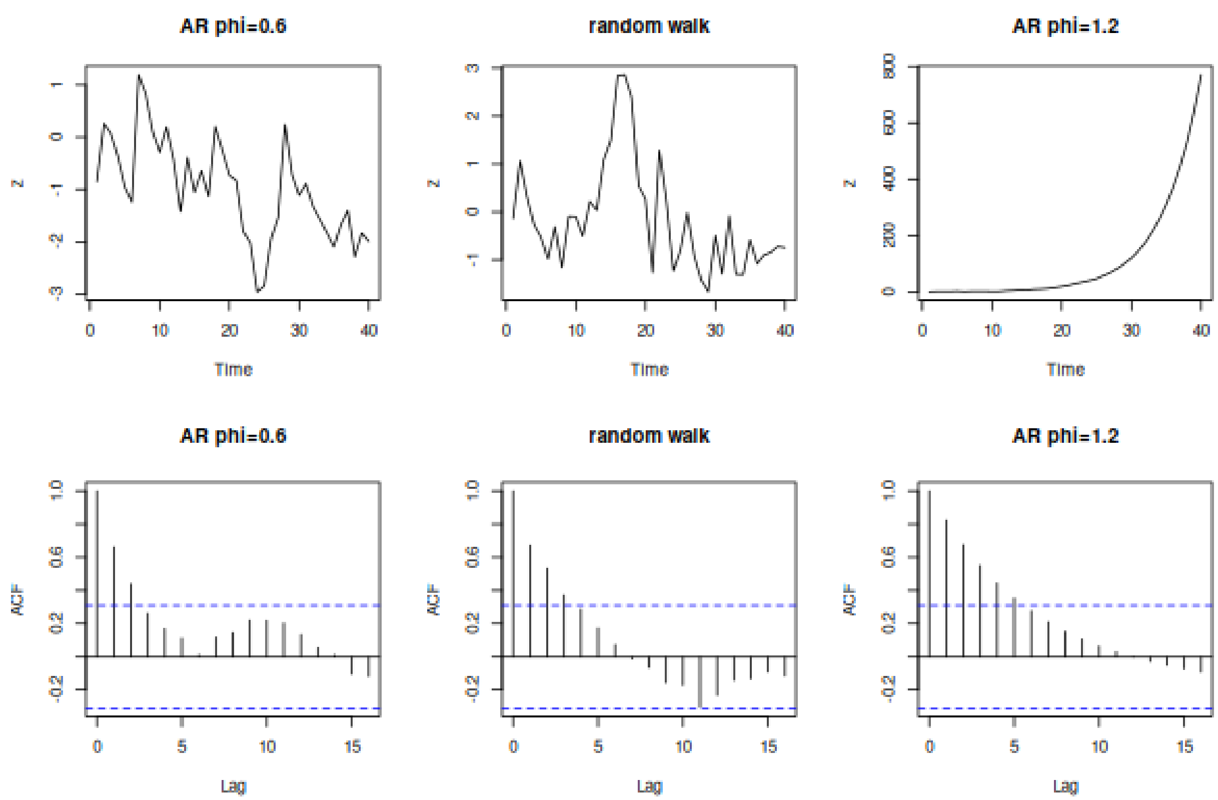
$$Y_{t+h} - \mu = \phi^h (Y_t - \mu) + \phi^{h-1}\epsilon_{t+1} + \cdots + \phi\epsilon_{t+h-1} + \epsilon_{t+h}, \quad h \geq 2.$$

Parameters $\mu, \phi, \sigma_\epsilon$ are estimated from the training data.
Forecasting: $\widehat{y}_{t+1|t} = \widehat{\mu} + \widehat{\phi}(y_t - \widehat{\mu})$ (regression towards the mean), $\widehat{y}_{t+h|t} = \widehat{\mu} + \widehat{\phi}^h(y_t - \widehat{\mu})$ and $\widehat{y}_{t+1|t} \to \widehat{\mu}$ as $h \to \infty$. For a dependent stationary process, there should be a better rule than the sample mean for short-term forecasts.
For $h = 2$, need $\mathrm{E}\,(Y_{t+2}|Y_1 = y_1, \ldots, Y_t = y_t) = \mathrm{E}\,(Y_{t+2}|Y_t = y_t)$ for AR(1).

(b) $\phi = 1$: $Y_{t+1} = Y_t + \epsilon_{t+1}$, this is a *random walk*. Note that $\text{Var}\,(Y_{t+1}) = \text{Var}\,(Y_t) + \sigma_\epsilon^2$, so if the process starts with non-random $y_0$, then $\text{Var}\,(Y_{t+1}) = (t+1)\sigma_\epsilon^2$ (exercise). That is, the variance is increasing linearly in $t$. Forecasting: $\hat{y}_{t+h|t} = y_t$ for $h > 0$.

(c) $|\phi| > 1$: The process is "exploding" to $\pm\infty$. Why is this clear from (1)?

Some plots of simulated time series generated from autoregressive processes with Gaussian (normally-distributed) innovations.



## Simple exponential smoothing

Exponential (positive) weighted average of past observations. For a constant $\theta \in (0,1)$, (note that sum of weights is 1)

$$\hat{y}_{t+1|t} = (1-\theta)y_t + (1-\theta)\theta y_{t-1} + (1-\theta)\theta^2 y_{t-2} + \cdots$$

Note that with an infinite past, from a geometric sum,

$$\sum_{i=0}^{\infty}(1-\theta)\theta^i = (1-\theta)\cdot(1-\theta)^{-1} = 1.$$

The stochastic model with an additive innovation (noise), from (*) on p 3, is

$$Y_{t+1} = (1-\theta)Y_t + (1-\theta)\theta Y_{t-1} + (1-\theta)\theta^2 Y_{t-2} + \cdots + \epsilon_{t+1}, \quad t = 1,2,\ldots$$

where $\epsilon_i$ are iid with mean 0 and variance $\sigma^2_\epsilon$ and the innovation $\epsilon_{t+1}$ is independent of $Y_1,\ldots,Y_t$. Then

$$\begin{aligned}
Y_t &= (1-\theta)Y_{t-1} + (1-\theta)\theta Y_{t-2} + (1-\theta)\theta^2 Y_{t-3} + \cdots + \epsilon_t \\
Y_{t+1} - \theta Y_t &= (1-\theta)Y_t + 0Y_{t-1} + 0Y_{t-2} + 0Y_{t-3} + \cdots + \epsilon_{t+1} - \theta\epsilon_t \\
Y_{t+1} &= Y_t + \epsilon_{t+1} - \theta\epsilon_t
\end{aligned}$$

The differenced series leads to a simpler representation; and partially explains why differencing is used in the Box-Jenkins methodology to get stationary series after differencing.

Recursion of simple exponential smoothing (usually written as):

$$\widehat{\ell}_t = \alpha y_t + (1-\alpha)\widehat{\ell}_{t-1}; \quad \widehat{y}_{t+1|t} = \widehat{\ell}_t, \quad t = 1, 2, \ldots$$

$\widehat{\ell}_t$ is a convex combination of the most recent observation and the previous smoothed value. $\widehat{\ell}_{t-1}$ is a geometric sum of $y_{t-1}, y_{t-2}, \ldots$. Hence

$$\begin{aligned}
\widehat{y}_{t+1|t} &= \alpha y_t + (1-\alpha)\widehat{\ell}_{t-1} \\
&= \alpha y_t + (1-\alpha)[\alpha y_{t-1} + (1-\alpha)\widehat{\ell}_{t-2}] \\
&= \alpha y_t + (1-\alpha)\alpha y_{t-1} + (1-\alpha)^2[\alpha y_{t-2} + (1-\alpha)\widehat{\ell}_{t-3}] \\
&\approx \alpha y_t + \alpha \sum_{i=1}^{t-1} (1-\alpha)^i y_{t-i} \\
\widehat{y}_{t+2|t} &= \alpha \widehat{y}_{t+1|t} + (1-\alpha)\widehat{\ell}_t = \widehat{\ell}_t \quad \text{(because } y_{t+1} \text{not known for 2-step forecast)} \\
\widehat{y}_{t+h|t} &= \widehat{\ell}_t, \quad t > 1.
\end{aligned}$$

The 1-step forecast is a geometric weighted average. Write the stochastic model (without hat on $\ell$) for the recursion as:

$$Y_{t+1} = L_t + \epsilon_{t+1}, \quad L_t = \alpha Y_t + (1-\alpha)L_{t-1} = L_{t-1} + \alpha(Y_t - L_{t-1}) = L_{t-1} + \alpha\epsilon_t$$

Then

$$\Delta Y_{t+1} := Y_{t+1} - Y_t = (L_t - L_{t-1}) + \epsilon_{t+1} - \epsilon_t = \alpha\epsilon_t + \epsilon_{t+1} - \epsilon_t = \epsilon_{t+1} - (1-\alpha)\epsilon_t$$

The previous $\theta$ matches $1 - \alpha$.

## Pseudo-code for rmse (simple exponential smoothing)

## Lab exercise: code and verify the output of R using parameter estimates from `HoltWinters()`

Part 1:

- Input `train` with size $n$.
- Estimate $\alpha$ parameter as $\widehat{\alpha}$ and get the smoothed series $\widehat{\ell}_2, \ldots, \widehat{\ell}_n$; $\widehat{\ell}_n$ is the last smoothed value of the training set.
- Output $\widehat{\alpha}$, $\widehat{\ell}_n$.

Part 2: Separate out-of-sample rmse from exponential smoothing

- Input $\widehat{\alpha}$, $\widehat{\ell}_n$, `holdout` with size $n_{holdout}$
- sse$\leftarrow$ 0
- fc$\leftarrow$ $\widehat{\ell}_n$; yt$\leftarrow$ holdout[1]; $\ell_{new} \leftarrow \widehat{\ell}_n$; fcvec[1]$\leftarrow$ $\widehat{\ell}_n$; fcerror$\leftarrow$ yt-fc; sse$\leftarrow$ sse$+$ fcerror$^2$.
- for i in $2, \ldots, n_{holdout}$:
- $\ell_{new} \leftarrow \widehat{\alpha}\times$ holdout[i-1] $+(1-\widehat{\alpha}) \times \ell_{new}$; fc$\leftarrow$ $\ell_{new}$ ; fcvec[i]$\leftarrow$ $\ell_{new}$ ; yt$\leftarrow$ holdout[i]; fcerror$\leftarrow$ yt-fc; sse$\leftarrow$ sse$+$ fcerror$^2$.
- end for
- return rmse=sqrt(sse/$n_{holdout}$) and fcvec (forecast vector)