Notes for Thursday, April 16, 1998

ARIMA Models

ARMA Models

A stationary autoregressive moving average process, $ARMA(p,q)$, satisfies *both* the equations for an $AR(p)$ process and a $MA(q)$ process:

$$\Phi_p(B)(y_t - \mu) \;=\; \Theta_q(B) z_t,$$

where

$$\begin{aligned} \Phi_p(B) &= 1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p \\ \Theta_q(B) &= 1 - \beta_1 B - \beta_2 B^2 - \ldots - \beta_q B^q \end{aligned}$$

where again z_t is a mean-zero white noise process [usually iid $N(0,\sigma^2)$], and $By_t = y_{t-1}$ is the *backshift operator.*

ARIMA Models

Many time series encountered in practice are non-stationary, so that the ARMA model described above is inadequate. However, the *differenced* series may well be stationary, as we have seen before. Let $\nabla = 1 - B$, so that $\nabla y_t = y_t - y_{t-1}$. Then an autoregressive integrated moving average process $ARIMA(p,d,q)$ is defined by the relation

$$\Phi_p(B)\nabla^d(y_t - \mu) \;=\; \Theta_q(B) z_t,$$

For example, for $d = 1$ the first differences $y_t - y_{t-1}$ is assumed to follow an $ARMA(p,q)$ model; and when $d = 2$, the second differences $y_t - 2y_{t-1} + y_{t-2}$ are assumed to be $ARMA(p,q)$. Note that an $ARIMA(p,0,q)$ model is just an $ARMA(p,q)$; $ARIMA(0,0,q)$ is $MA(q)$, etc.

The main tool for fitting $ARIMA(p,d,q)$ models in SPLUS is the function `arima.mle`; residual plots and goodness-of-fit tests are available with `arima.diag`.

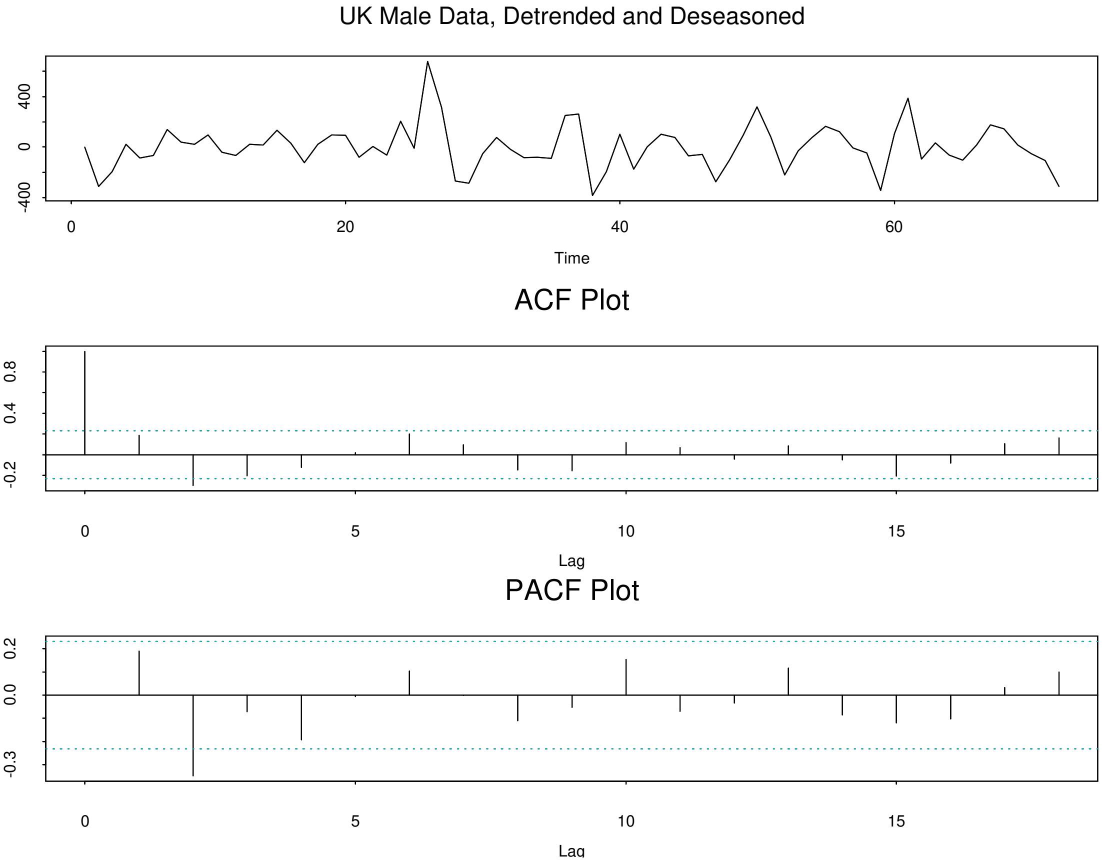
Example: UK male respiratory disease deaths

After taking out some cyclic structure. Recall from our discussion of spectral analysis (or just by looking at a time series plot directly!) that there was a clear 12-month cycle in the `uk.male` time series. We may take this cycle out of the data, along with a linear secular trend, as follows, and then examine the residuals for time series modeling:

```
402 > uk.reg _ lm( y ~ time + osc1 + osc2,
+ data.frame(y=uk.male, time=1:72,
+ osc1=cos(2*pi*(1:72)/12), osc2=sin(2*pi*(1:72)/12)) )
402 > uk.deseasoned _ residuals(uk.reg)
402 > uk.deseasoned _ cts(uk.deseasoned,start=dates("01/01/74"),
```

```
+ units="months")
402 > par(mfrow=c(3,1))
402 > ts.plot(uk.deseasoned)
402 > acf.plot(acf(c(uk.deseasoned),plot=F),main="ACF Plot")
402 > acf.plot(acf(c(uk.deseasoned),plot=F,type="p"),main="PACF Plot")
```

The `arima.mle()` function has an argument `xreg=` that can automatically regress out trends, seasonal components, and other nonrandom structure in the time series, but it is more informative, at least for exploratory purposes, to regress them out by hand and then fit time series models to the residuals as we will do.



It appears from the ACF plot that an MA(2) or so will fit; from the PACF it looks like an AR(2) or so might do also. We will try several and see where we get.

First, some MA(q) models, which we will compare using the Akakike Information Criterion (AIC). AIC $= -2 \cdot$ (maximized log-likelihood) $+ 2 \cdot$ (# of parameters), i.e. the log-likelihood chi-squared statistic penalized by the number of parameters in the model.

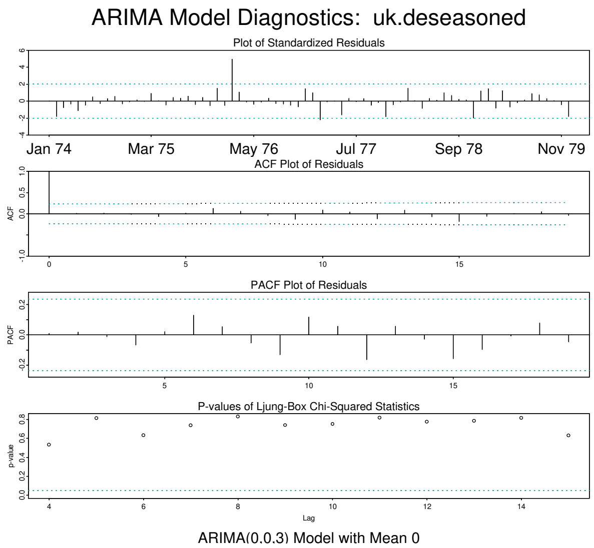
```
402 > (ma1 _ arima.mle(uk.deseasoned,model=list(order=c(0,0,1))))$aic
[1] 943.2947
402 > (ma2 _ arima.mle(uk.deseasoned,model=list(order=c(0,0,2))))$aic
[1] 940.1631
402 > (ma3 _ arima.mle(uk.deseasoned,model=list(order=c(0,0,3))))$aic
[1] 938.8625
```

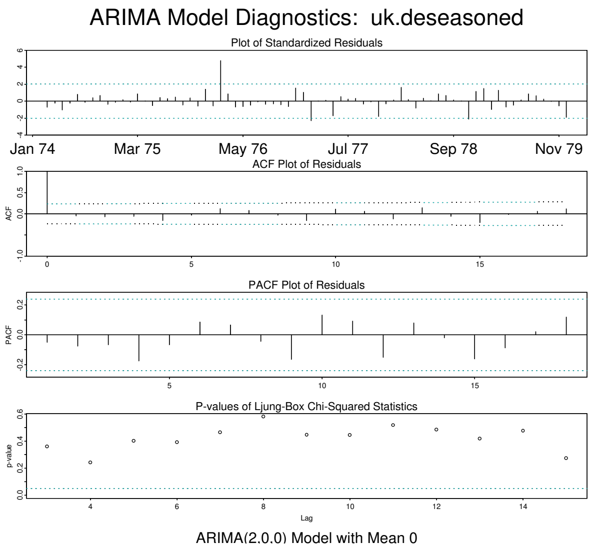
```
402 > (ma4 _ arima.mle(uk.deseasoned,model=list(order=c(0,0,4))))$aic
[1] 940.7722
402 > (ma5 _ arima.mle(uk.deseasoned,model=list(order=c(0,0,5))))$aic
[1] 942.7262
```

Using the AIC criterion, the best (minimum AIC) model appears to be MA(3).

For AR(p) processes, comparison of models by AIC can be done automatically with the `ar()` function. For comparison, we find the best AR(p) model with `ar()` and then examine residual plots produced by `arima.diag()`:

```
402 > ar(uk.deseasoned)$order  # see help(ar)
[1] 2
402 > (ar2 _ arima.mle(uk.deseasoned,model=list(order=c(2,0,0))))$aic
[1] 911.467
402 > arima.diag(ma3)
402 > arima.diag(ar2)
```





The plots on the left are for the MA(3) model, and the ones on the right are for the AR(2) model. The residual plots show estimates of z_t for the two models and are almost self-explanatory; as usual we want to see evidence that the z_t are approximately iid $N(0, \sigma^2)$. The last plot shows a sequence of p-values for the so-called "Portmanteau" test statistics, based on the estimated autocovariance $\hat{\gamma}(k)$:

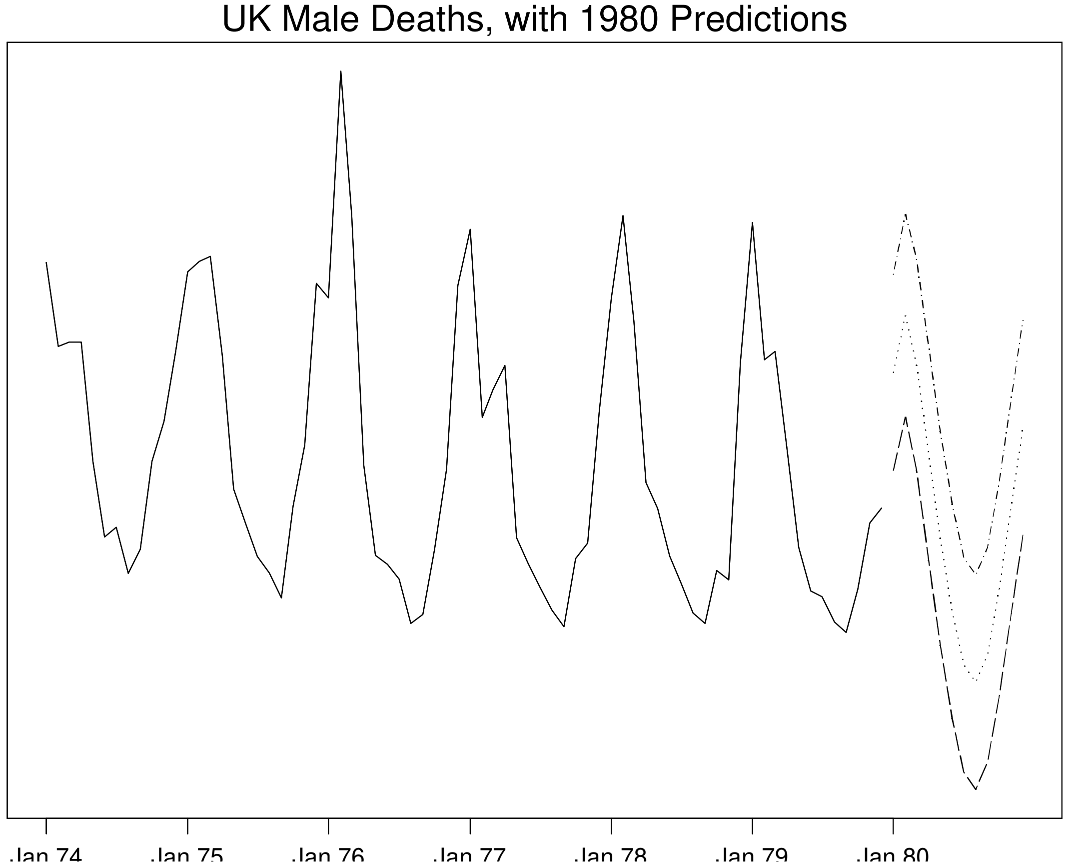
$$Q_K = n \sum_{k=1}^{K} \hat{\gamma}(k)\ \chi^2_{K-r}, \text{ approximately,}$$

when the right $ARIMA(p, d, q)$ model is fitted, with $r = p + q$. Q_K is plotted for several values of K, in practice the p-values for K between 10 and 20 are worth looking at. There is no evidence of lack of fit for either of these models; the only troubling aspect is one or two large residuals in *both* residual plots, especially just preceding Jan 1976. Based on the lower AIC criterion, we would probably choose the AR(2) model here.

Finally we illustrate forecasting using the `arima.forecast` function.

```
402 > uk.fore _ arima.forecast(uk.deseasoned,n=12,model=ar2$model)
402 > uk.fore$mean _ uk.fore$mean + predict(uk.reg,data.frame(time=73:84,
+ osc1=cos(2*pi*(73:84)/12), osc2=sin(2*pi*(73:84)/12)) )
402 > uk.fore$mean _ cts(uk.fore$mean,start=dates("01/01/80"),units="months")
402 > uk.fore$stderr _ cts(uk.fore$stderr,start=dates("01/01/80"),units="months")
402 > ts.plot(uk.male,uk.fore$mean, uk.fore$mean+2*uk.fore$std.err,
+ uk.fore$mean-2*uk.fore$std.err)
402 > title("UK Male Deaths, with 1980 Predictions")
```

The resulting plot reflects uncertainty due to predicting the ARIMA(2,0,0) process itself, but not uncertainty due to estimating parameters for the trend and seasonal component and the AR(2) model itself.



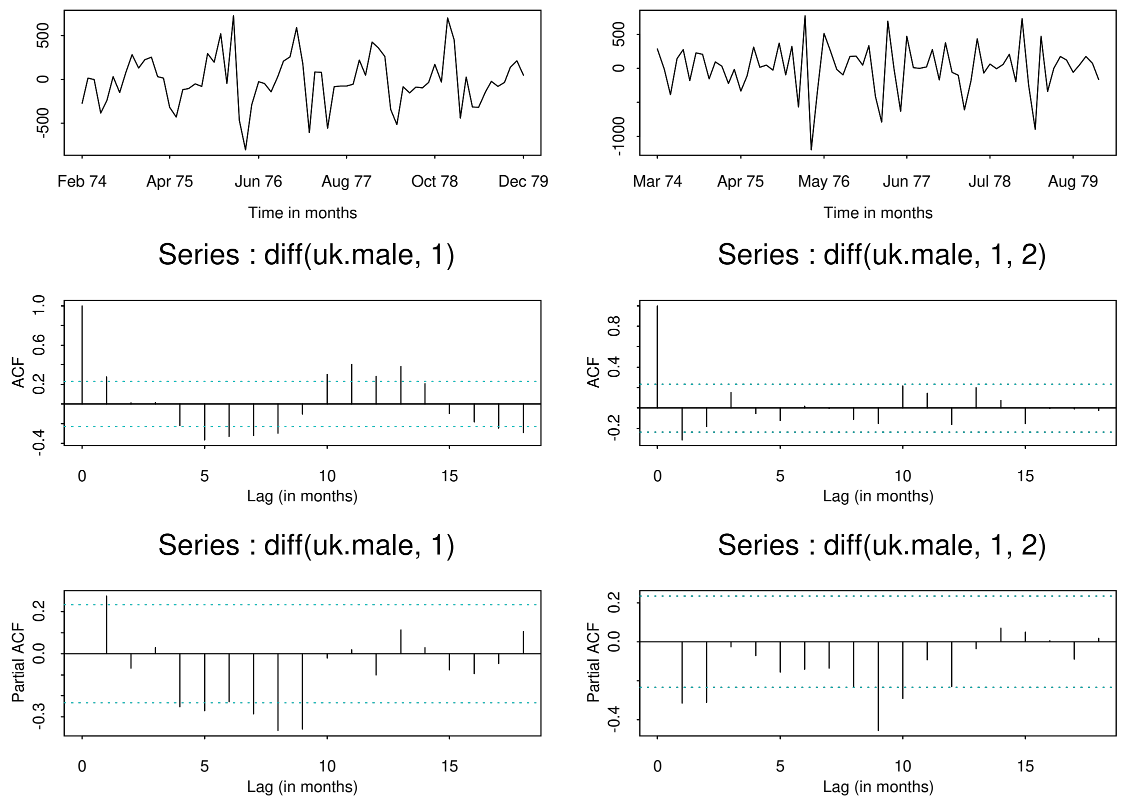
Differencing. Looking at the time series plot again, we see that the peak preceding Jan 1976 is causing the outliers in our earlier data.

We can get rid of most of the anomalous peak by taking first or second differences of the original series.

```
402 >  par(mfcol=c(3,2))
402 >  ts.plot(diff(uk.male,1))
402 > invisible(acf(diff(uk.male,1)))
402 >  invisible(acf(diff(uk.male,1),type="p"))
```

```
402 >  ts.plot(diff(uk.male,1,2))
402 >  invisible(acf(diff(uk.male,1,2)))
402 >  invisible(acf(diff(uk.male,1,2),type="p"))
```

On the left, we plot $\nabla y_t = y_t - y_{t-1}$ and its ACF and PACF plots; on the right we plot $\nabla^2 y_t = y_t - 2y_{t-1} + y_{t-2}$ and its ACF and PACF plots.



It looks as though the second differences $\nabla^2 y_t$ may be promising for MA(q) modeling. We also go ahead and search for an AR(p) model.

```
402 > (arima011 _ arima.mle(uk.male,model=list(order=c(0,1,1))))$aic
[1] 1003.314
402 > (arima012 _ arima.mle(uk.male,model=list(order=c(0,1,2))))$aic
[1] 1005.262
402 > ar(diff(uk.male,1,2))$order
[1] 12
402 > (arima12.1.0 _ arima.mle(uk.male,model=list(order=c(12,1,0))))$aic
[1] 803.6236
```

The MA models quickly confirm what we see in the ACF plot, that an MA(1) model fits better than others. By inspecting the `arima011$model` component, we may write this model in the ARIMA notation, $\Phi_0(B)\nabla^2 y_t = \Theta_1(B)z_t$ as

$$y_t - 2y_{t-1} + y_{t-1} = (1 - (-0.3089807)B)z_t = z_t + (0.3089807)z_{t-1}$$

The AR model selected has twelve components (can you see this in the PACF plot also?), and would be written as

$$\Phi_{12}(D)\nabla^2 y_t = \Theta_0(D)z_t = z_t$$

where again the coefficients of the polynomial $\Phi_{12}(D)$ could be read off the `arima12.1.0$model` component. This AR fit suggests that there is still a 12-month structure in the data. However the `arima.diag` plots (omitted here) suggest that, although the residuals are uncorrelated, the models do not fit very well (e.g. P-values for the Q_K test statistic handily reject both models).

Taking small differences does not seem to be removing the seasonality in the data.

Seasonal Models

Venables and Ripley (throughout Chapter 15, but especially pp. 451–453) also consider the `uk.male` data, and they illustrate another approach to removing the seasonality.

Since there is clearly a 12-month cycle in the data, it makes sense to consider not

$$\nabla y_t = (1 - B)y_t = y_t - y_{t-1}$$

but rather

$$\nabla_{(12)} = (1 - B^{12})y_t = y_t - y_{t-12}$$

In other words, there is a deterministic 12-month cycle that is always the same, but the year-to-year differences we will model as a random ARMA(p, q) process. This leads to considering models of the form

$$\Phi_P(B^{12})\nabla^D_{(12)} y_t = \Theta_Q(B^{12})z_t$$

which we would write as ARIMA$(P, D, Q)_{12}$, indicating an ARIMA model, but with period 12 (in general of course, 12 could be replaced with any period s that makes sense for the problem). This can be combined with terms that operate on B directly, for example

$$\Phi_P(B^{12})\phi_p(B)\nabla^D_{(12)}\nabla^d y_t = \Theta_Q(B^{12})\theta_q(B)z_t$$

which would be written ARIMA$(p, d, q) \times (P, D, Q)_{12}$.

Obviously there is a forest, or even a jungle, of models to consider here. By working slowing up from simple models, one can arrive at fairly sensible results; Venables and Ripley's analysis on pp. 451–452 is a nice example of this. Two of their final models are illustrated below (see the book to see the path that got them there):

- One of their final models is an ARIMA$(2, 0, 0) \times (2, 0, 0)_{12}$:

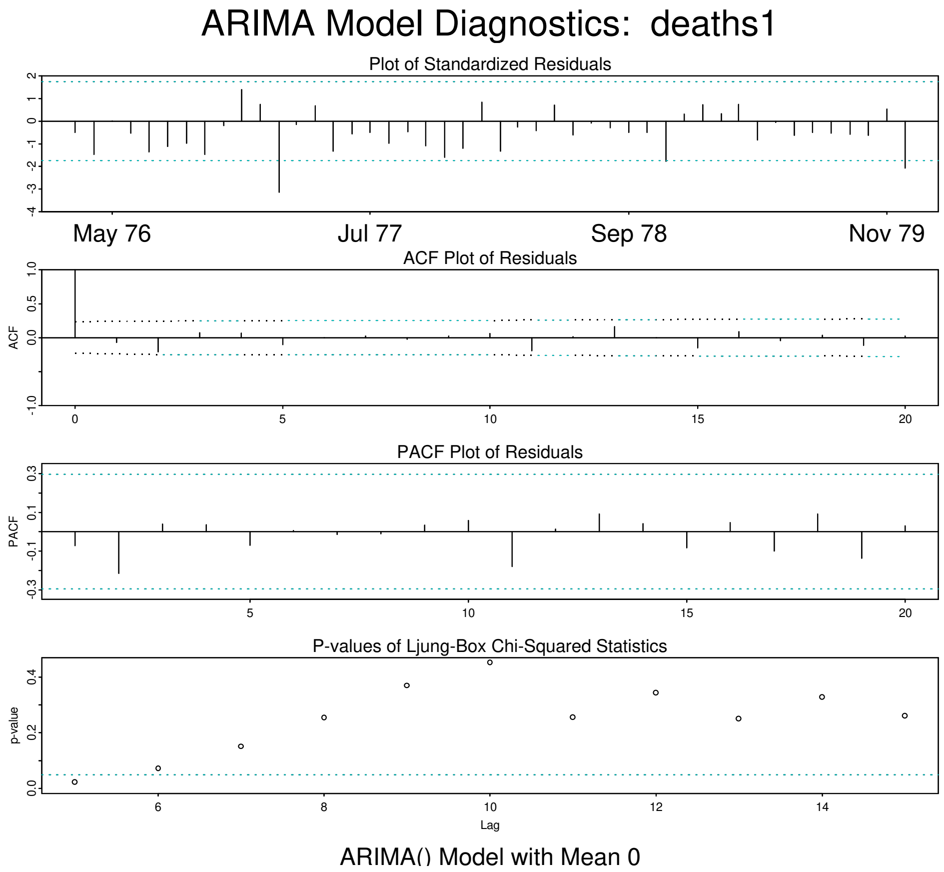
```
402 > deaths _ uk.female + uk.male
402 > deaths1 _ deaths - mean(deaths)
402 > deaths.arima4 _ arima.mle(deaths1,model=list(
+  list(order=c(2,0,0)), list(order=c(2,0,0),period=12)) )
402 > deaths.arima4$aic
[1] 634.0667
402 > deaths.arima4$model[[1]]$ar
[1]  0.4782125 -0.2787332
```

```
402 > deaths.arima4$model[[2]]$ar
[1] 0.1734640 0.6347404
402 > arima.diag(deaths.arima4)
```

This produces the model

$$\Phi_2(B^{12})\phi_2(B)\nabla^0_{(12)}\nabla^0 y_t \;=\; \Theta_0(B^{12})\theta_0(B)z_t$$
$$(1-(0.173464)B^{12})-(0.634740)B^{24})(1-(0.4782125)B+(0.2787332)B^2)(y_t-\mu)=z_t$$

with diagnostic plots



- The last of their models is an $\text{ARIMA}(1,1,0)_{12}$:

```
402 > deaths.arima5 _ arima.mle(deaths1,model=list(
+ order=c(1,1,0),period=12))
402 > deaths.arima5$aic
[1] 691.2965
402 > deaths.arima5$model
$order:
[1] 1 1 0

$period:
[1] 12
```

```
$ar:
[1] -0.6057538

$ndiff:
[1] 1
```

This yields the model

$$\begin{aligned} \Phi_1(B^{12})\nabla^1_{(12)}y_t &= \Theta_0(B^{12})z_t \\ (1+0.6057538B^{12})(1-B^{12})y_t &= z_t \end{aligned}$$

which is not quite the same as Venables and Ripley (They got 0.727 instead of 0.606; I'm not sure why at the moment...).

The diagnostic plot follows; in general this looks like a less-good fit to me that `deaths.arima4` did. Note that this model has the structure of a constrained ARIMA$(2,0,0)_{12}$, which might have provided a better fit.