Homework 4 solutions

Joe Neeman

October 27, 2010

1. We began by looking at the ACF of the original data sequence (Figure 1), which seems to decay very slowly. In particular, the process is probably not an ARMA process. The ACF and PACF of the first differences (Figure 1) look much more plausible. If the first differences were an AR or an MA process, we would expect either the ACF or the PACF to cut off after a finite lag. Since this doesn't seem to happen, we will propose an ARMA model; the simplest candidate is ARMA(1,1), so let's start with that.

   Using `R`'s `arima` function, we estimated the AR coefficient to be 0.27 (with a standard error of 0.11) and the MA coefficient to be -0.8180 (with a standard error of 0.06). The mean (of the differenced sequence, so it corresponds to the drift of the original sequence) was estimated at 0.006 with a standard error of 0.003. Some diagnostic plots can be seen in Figure 2. The correlations of the residuals were not significant, although some of were fairly close. The Shapiro-Wilk test returned a fairly reasonable p-value of 0.64 and the QQ-plot suggests the the residuals are approximately normal.

   Nevertheless, we could attempt to reduce the almost-significant correlations at lags 2 and 4 by introducing some more MA terms. In fact, doing so reduces the AIC of the fitted model. After trying MA degrees of 1 through

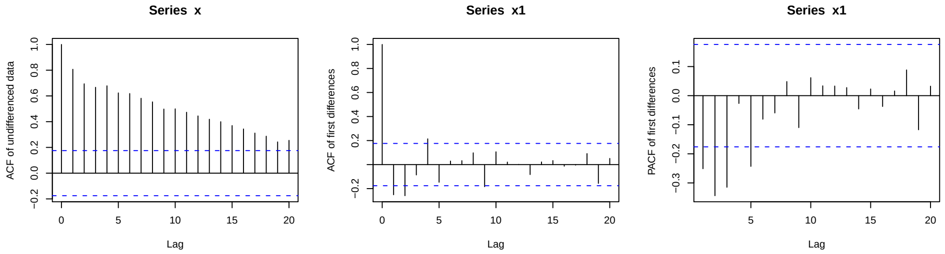

Figure 1: ACF of the undifferenced data and the first differences, and PACF of the first differences.

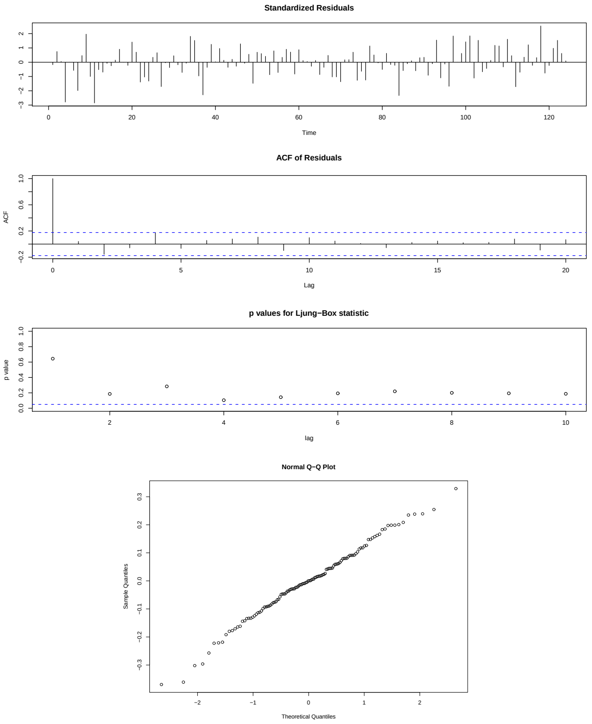

Figure 2: Diagnostic plots for the ARMA$(1, 1)$ model on the differenced data.

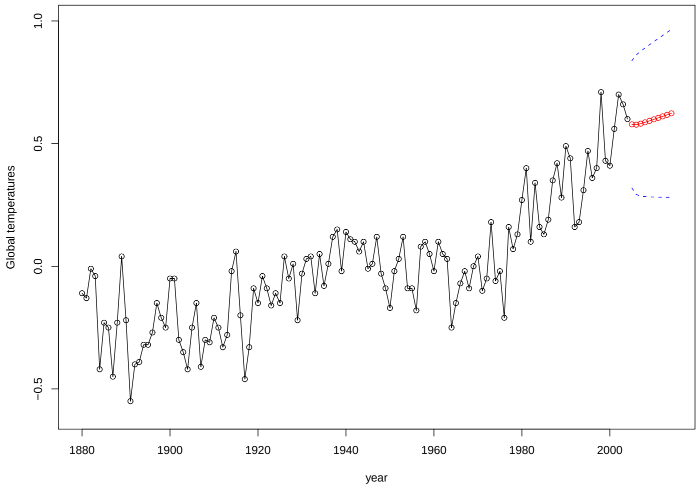

Figure 3: Predictions from the ARIMA(1, 1, 1) model.

5, we found that an ARMA(1, 4) model had the lowest AIC (-149.52). It's Shapiro-Wilk p-value was 0.82 and the other diagnostic plots produced by `tsdiag` seemed reasonable, but its BIC was actually higher than that of the ARMA(1, 1) model. In fact, the ARMA(1, 1) model had the lowest BIC of all the alternatives we tried. Since simpler models are nice for various reasons, we decided to stick with ARMA(1, 1) for the predictions (ie. ARIMA(1, 1, 1) once we undifference the sequence).

The predictions are shown in Figure 3. An upward drift is evident in the predictions (corresponding to the mean of 0.06 that we noticed in the differenced data), but the standard error is quite large.

2. Once again, we began by looking at the ACF of the original data. Since it decayed slowly, we took first differences and looked at the ACF again. This time, there were values at lags of multiples of 12 which decayed slowly, so we took 12th differences and looked at both the ACF and the PACF. All of these plots are in Figure 4.

   Let us look at the seasonal component first: in the ACF of the final differenced series, there is a strong correlation at lag 12, but no strong correlations at larger multiples of 12. This suggests an MA(1) model for the seasonal part. We could also try fitting an AR(3) model, since the PACF has peaks at lags of 12, 24 and 36.

   For the non-seasonal component, an AR(2) model looks likely because the PACF drops off dramatically after a lag of two. An MA(4) model is also possible, but it looks less compelling. In any case, we fit four models to

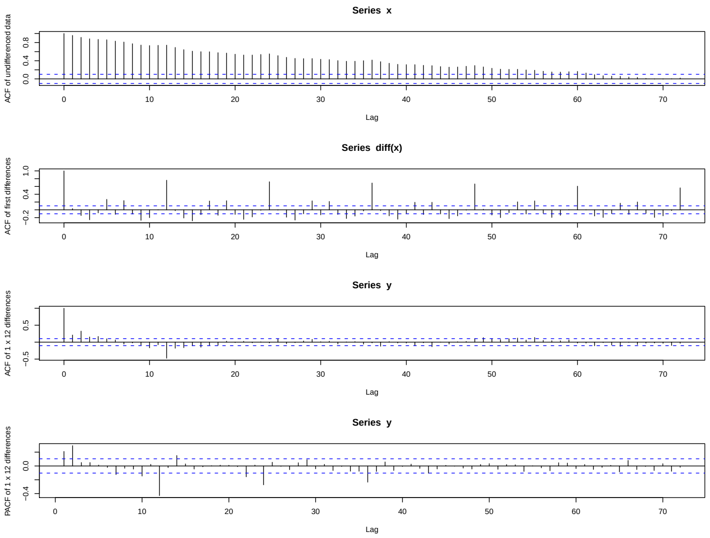

Figure 4: ACF and PACF of original and differenced sequences.

our differenced data using all of the possible combinations that we just mentioned. In terms of AIC, the winner was $(2,0,0) \times (3,0,0)_{12}$, with an AIC of 3228.16. In close second place was $(2,0,0) \times (0,0,1)_{12}$, with an AIC of 3229.77. As in the previous question, the BIC gave the reverse picture, preferring the simpler model (with a score of 6.158) to the more complex one (6.172). We also checked tried combining the seasonal parts of the two models with a $(2,0,0) \times (3,0,1)_{12}$ model, but it was not the preferred model under either AIC or BIC.

Our two main contenders so far are attempting to estimate the mean, but in both cases it does not seem to be significant (-0.11 ± 0.59 or -0.20 ± 0.72). Therefore, we fit a model without a mean instead. This reduced the AIC in both cases, although it did not change the relationship between them.

As we did in the previous question, we will go with the simpler model, $(2,0,0) \times (0,0,1)_{12}$, for our predictions. First, though, let's make sure that the diagnostics are reasonable. Some diagnostic plots are in Figure 5. Note that we have a single outlier near the end of the sequence. Indeed, the Shapiro-Wilk test reports a p-value of 0.00007, but it jumps to 0.38 if we remove that single outlier! The Q-Q plot tells a similar story, so the model does seem a reasonable fit.

Remember that everything so far was fitted to the differenced data. When we fit it to the original data, we will fit a $(2,1,0) \times (0,1,1)_{12}$ model (without a drift term). The resulting predictions are given in Figure 6. As in the

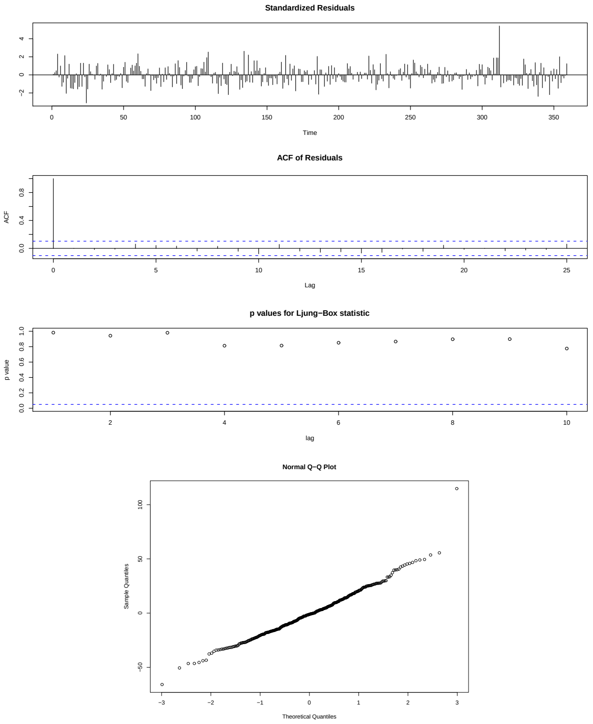

Figure 5: Diagnostics for $(2, 0, 0) \times (0, 0, 1)_{12}$ (on the differenced series).

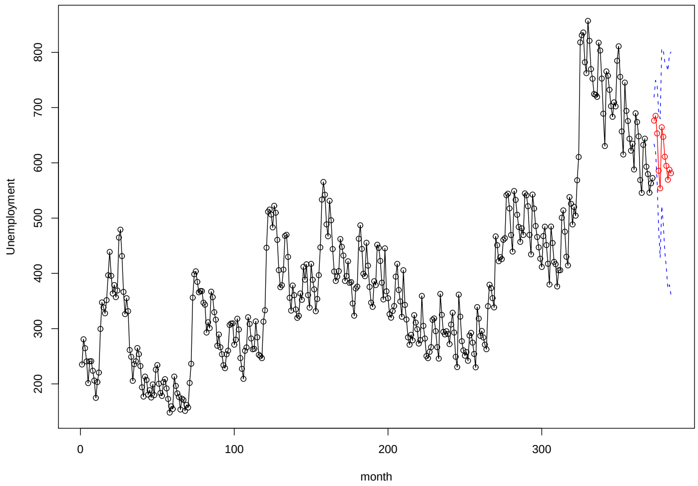

Figure 6: Predictions from a $(2,1,0) \times (0,1,1)_{12}$ model.

previous question, the error bars are quite wide: the estimated variance was 450.