Handling and Finding

Shuoyang Shi

12/22/2020

Data 1: (univariate) "Carbon dioxide emissions in Hawaii.csv"

(S)ARIMA approach

Looking at the plot, the data seem to be non-stationary. This is also confirmed by the acf plot. Huge spikes all over the places. After trying detrending and differencing. We go for differencing, because the detrended series seem to be non-stationary. But after the first differencing we still see spikes at seasonal levels (lags that are multiple of 12). This indicates that we should also consider first seasonal differencing. The resulting series seems now stationary: at nonseasonal level, its acf cutting off after lag 1 and pacf cutting off after lag 1 suggesting an ARMA(1,1).

Here I will try five models and compare their AIC and BIC. The AIC prefer the ARIMA(1,1,1)(0,1,1)[12], whereas the BIC prefers the simpler ARIMA(0,1,1)(0,1,1)[12] model. It is often the case that the BIC will select a model of smaller order than the AIC or AICc.

I would say ARIMA(1,1,1)(0,1,1)[12] is better after comparing the diagnostic check. The diagnostic checking looks very good: Errors seem to be random without a specific pattern, QQ plot reasonably indicates that normality is respected, no spikes in the acf of the errors indicating they are white, and also the Portmanteau test shows no dependence, and all P values are well above the 5% level. Hence, we can accept this as a final model and the estimated model is:

$$(1 - 0.2526B)(1 - B)(1 - B^{12})x_t = (1 - 0.5953B)(1 - 0.8568B^{12})w_t,$$
$$\text{with } \hat{\sigma}_w = 0.2826$$

The forecast for the next four months is: 354.8905, 355.6698, 356.5030 and 357.8046.

Unit root test

The Dickey-Fuller test confirms that the data has a unit root (P-value=.4)

Spectral analysis

There are two other peaks: one at 1 simply referring at the seasonal nature of the data (as they are monthly). Another small one at 2 which indicates a possible 1/2=0.5-year cycle in the Carbon dioxide emissions in Hawaii. Two methods agree the same result.

Then I get the model named fit1, which is:

$$X_t = 332.1623 - 1.6060*\cos(2\pi n)*0.083 + 1.7392*\sin(2\pi n)*0.083 + 0.8543*\cos(2\pi n)*0.167 - 0.166*\sin(2\pi n)*0.167$$

After comparing AIC, I conclude ARIMA(1,1,1)(0,1,1)[12] is the best model I could get.

Data 2: (univariate) "manufacturers-index-of-new-order.xlsx"

(S)ARIMA approach

Looking at the plot, the data seem to be non-stationary. This is also confirmed by the acf plot. Huge spikes all over the places. After trying detrending and differencing. We go for differencing, because the detrended series seem to be non-stationary. The first difference series seems now stationary: its acf cutting off after lag 2 and pacf cutting off after lag 2 suggesting an ARMA(2,2).

Here I will try four models and compare their AIC and BIC. The AIC prefer the ARIMA(2,1,1) fit, whereas the BIC prefers the simpler ARIMA(1,1,1) model. It is often the case that the BIC will select a model of smaller order than the AIC or AICc.

I would say ARIMA(2,1,1) is better after comparing the diagnostic check. The diagnostic checking looks very good: Errors seem to be random without a specific pattern, QQ plot reasonably indicates that normality is respected, no spikes in the acf of the errors indicating they are white, and also the Portmanteau test shows no dependence, and all P values are well above the 5% level. Hence, we can accept this as a final model and the estimated model is:

$$(1 + 0.2298B + 0.24B^2)(1 - B)x_t = -0.0878 + (1 + 0.1186B)\, w_t,$$
$$\text{with } \hat{\sigma}_w = 12.66$$

The forecast for the next four months is: 124.3165, 130.1649, 127.4162 and 126.5153.

Unit root test

The Dickey-Fuller test confirms that the data has a unit root (P-value=.48)

Spectral analysis

Method 2 is hard to find the result. From method 1, we could get there are 11-year and 5.5-year cycles in the Manufacturers' Index of New Orders of Durable Goods for United States.

Then I get the model named fit1, which is:

Xt = 85.295 + 33.154*cos(2πn)* 0.008 - 21.042*sin(2πn)* 0.008
+ 7.197* cos(2πn)* 0.015 + 20.119 *sin(2πn)* 0.015

After comparing AIC, I conclude ARIMA(2,1,1) is the best model I could get.

Data 3: (multivariate) "economic-indicators-time-series.xlsx"

Linear regression

Looking at the scatter plot, there seems convex, so first I center the public and then square it. I try two linear model one with public^2, one without public^2. AIC shows fit2 (private ~ trend + pub + pub2) is better. However, the residual standard error is too large, which means linear regression model is really bad.

Lagged regression

I use three methods here.

This first one is looking the lag2.plot of public and private. It shows fairly strong linear relationships between Private, PRIt, and the Public series at many lags, such as PUBt-7, PUBt-8, PUBt-15, PUBt-16, PUBt-17, PUBt-19, etc. Indicating the coefficients are negative, implying that increases in the Public lead to decreases in the Private. However, some plots have very dispersive points. In order to find a good fit, I would try PUBt-18,

PUBt-19, PUBt-30 and PUBt-31 separately, which have more agminated points. And then I tried fit7 (`private ~ L(public,18) + L(public,19) + L(public,30) + L(public,31)`) which is an ensemble method I learnt from machine learning. As a result, fit7 indeed is the best model among fit3 to fit 7.

The second method is looking at acf and pacf. Observations separated by six months are negatively correlated, showing that positive excursions tend to be associated with negative excursions six months removed. however, shows some departure from the cyclic component of each series and there is an obvious peak at h = −6, h = -18, h = -30, h = -42. This result implies that public measured at time t−6, t-18, t-30, t-42 months are associated with the private series at time t. So, I get fit8 (`private ~ L(public,6) + L(public,18) + L(public,30) + L(public,42)`). Which is better than fit7. Then using ensemble method again, use fit7 and fit8 together to get fit9 (`private ~ L(public,6) + L(public,18) + L(public,30) + L(public,42) + L(public,19) + L(public,30) + L(public,31)`), and fit9 is the best one for now.

The third method is using `LagReg()` , I also examine the inverse relation. MSE of the `public(t) = 29138.43 + -0.122755* private(t+5)` is smaller. Compared with fit9, I conclude fit 9 is the best one. So, my final model would be:

$$PRI_t = 101100 + 3.46*PUB_{t-6} - 3.426*PUB_{t-18} + -3.803*PUB_{t-30}$$

$$+ 2.471*PUB_{t-42} - 1.242*PUB_{t-19} + 0.6106*PUB_{t-31} + w_t .$$

The forecast of private for the next four months is: 62433.334, 59685.833, 58565.872 and 61892.549.

Coherence analysis

We see from the coherence graph between public and private data that there is a high dependence between the two series especially at the peak frequencies of 0.01, 0.08, 0.18, etc. The significance level for a positive coherence is computed to be C= 0.6479. Because there are so many peaks. It is hard for me to consider running a lagged frequency lagged regression according this plot.

Data 4: (multivariate) "wheatherPr.xlsx"

Linear regression

First, I transform Precip to the square of it, because after the transformation, the relationship between Temp and sqrt(Precip) is more linear. And then I have two models,

fit1 without sqrt(Precip) and fit2 with sqrt(Precip). After comparing AIC, I find fit2 is a better model for now.

Lagged regression

I use the lag2.plot for each pair of the variables. It shows fairly strong linear relationships between Temp(Tt) and the DewPt, CldCvr, WndSpd, sqrt(Precip) series at some lags, such as DPt-38, CCt-1 , WSt-9, P-8. After comparing fit3(`temp ~ L(dp,38) + L(cc,1) + L(ws,9) + L(P,8)`) and fit4(`temp ~ L(dp,38) + L(cc,1) + L(P,8)`), I get fit4 is better, however fit2 is better overall. However, fit2 cannot be used to do prediction. And the residual standard error of fit4 is really small too. Hence, my final model would be fit4:

TEMPt = 17.76701 + 1.08722 * DPt-38 -14.70077 * CCt-1 - 0.16402 * P-8

The forecast of temp for the next four months is: 22.222, 19.546, 17.991 and 11.781.

Coherence analysis

We see from the coherence graphs between each pair of variables there are too many peak frequencies. The significance level for a positive coherence is computed to be C= 0.38. It is hard for me to consider running a lagged frequency lagged regression according this plot.

Data 5: (multivariate) "NZBirths.csv"

Linear regression

First, the ccatterplot matrix showing relations between each pair of variables. All relationships are positive and fairly linear. So, I perform fit1 (`tm ~ mm + mf + tf`). But it is not good with residual standard error of 104.

Lagged regression

I use `LagReg()` here. And tried a several thresholds and got several different model. After comparing MSE between them, I conclude a final model which is

$$TM_t = 509.37542 + 1.62727*MMt - 0.01182*MMt\text{-}1 - 0.4447*MFt -$$

$$0.08*MFt\text{-}1 - 0.1261*MFt\text{-}2 + TFt + wt$$

The forecast of total male for the next four months is: 7493.124, 8352.738, 6723.981 and 7383.52.

Coherence analysis

We see from the coherence graph between each pair of data. For the graph of TM and MM, there is a high dependence between the two series especially at the peak frequency of .05. The significance level for a positive coherence is computed (similarly as done in class) to be C=.38. It is hard for me to consider running a lagged frequency lagged regression according this plot.

Data 6: (multivariate) "pub-prinv.xlsx"

Linear regression

From the scatter plot, I find there seems a positive linear relationship between prinv and govinv. Hence, I fit this with the simplest formula fit1 (`govinv ~ prinv`) with residual standard error 61.86 for this part.

Lagged regression and TFM

For this part, I will use three methods.

The first method is using `LagReg()`, and then find fit2 (`govinv ~ prinv + L(prinv,1) + L(prinv,2) + L(prinv,3) + L(prinv,5) + L(prinv,8)`)), however in the summary, it shows only `L(prinv,8)` is significant. Then I find fit3 (`govinv ~ prinv + L(prinv,8)`), whose residual standard error is 43.63. It is the best model for now.

Then, second method is using `lag2.plot`, due to method 1, we focus on prinv(t-8), there are two slopes we need to find out (when x>400 or <400). So I use dummy variable: fit5(`govinv ~ prinv + priL8*dL8, data=inv, na.action=NULL`), whose residual standard error 35.52 is smallest for now.

Moreover, I perform TFM. After looking at the acf and pacf of detrended prinv, acf dying down and pacf cutting of after lag 1. I suggest the detrended prinv an AR(1) model with the ar1 coefficient of `0.8912` . Then get the prewhitened detrended prinv series, and the filtered govinv series. Sample CCF of the prewhitened, detrended prinv and the similarly transformed govinv series; negative lags indicate that prinv leads govinv. Noting the apparent shift of d = 8 months and the decrease thereafter, it seems plausible to hypothesize a model of the form:

$$\alpha(B) = \frac{\delta_0 B^8}{1 - \omega_1 B}$$

$$\therefore \quad y_t = \alpha(B)\, x_t + \eta_t$$

$$= \frac{\delta_0 B^8}{1 - w_1 B}\, x_t + \eta_t$$

$$\Rightarrow \quad y_t = \alpha + w_1 y_{t-1} + \delta_0 x_{t-8} + u_t$$

we could get coefficients by R.

Hence, the model would be:

govinv → y_t ; prinv → x_{t-8}

Final Model.

$$y_t = 8.62 + 0.99\, y_{t-1} + 0.059\, x_{t-8} + u_t,$$

where $u_t = 0.47\, u_{t-1} + w_t$,

where w_t is the white noise with

$\sigma_w^2 = 57.83$

AIC of this model is 6.96, which is smaller than the AIC of fit 5.

Hence, the model by using TFM is the best model I could get.

The prediction would be: 812.71, 849.57, 886.05 and 922.17.

And the final model is great. According to the diagnostic check. The diagnostic checking looks very good: Errors seem to be random without a specific pattern, QQ plot reasonably indicates that normality is respected, no spikes in the acf of the errors indicating they are white, and also the Portmanteau test shows no dependence, and all P values are well above the 5% level.

Coherence analysis

We see from the coherence graph between prinv and govinv data that there are many high dependences between the two series, because I can see many frequency peaks. The significance level for a positive coherence is computed to be C= 0.65. It is hard for me to consider running a lagged frequency lagged regression according this plot.