**Stat 510 Week 13 Solutions**

1. Complete problem 5.1 on page 315 of the textbook. The data are in the file fracdiff.dat in the Week 13 folder.

Here are a few pointers and suggestions -

Use the code in Lesson 13.1 for guidance.

First, you’ll have to install the fracdiff package. Either in the R menus, use Install Packages and select fracdiff or use the command install.packages("fracdiff"). After the installation, be sure to then use the command library(fracdiff).

For part c, the command is fracdiff(y, nar=1, nma = 0, M = 30) where y would be the name of the series that you’re analyzing.. The estimates will be given in the output so report them and their significance as the answer. You may compute the test statistic by hand; fracdiff(x,nar=1,nma=0,M=30)$stderror.dpq will provide the standard errors. Or you may use the summary command and R will print the test statistics and corresponding p-values:

mymodel <- fracdiff(x, nar=1, nma=0, M=30)

summary(mymodel)

The M = 30 parameter of the command is a bit mysterious. It has to do with the setup of the maximum likelihood estimation of the parameters. It’s what the authors use so we’ll use it too.

In question 2 of the homework, we’ll look at how we would identify this particularly fractionally differenced model in practice.

For parts e and f, you’ll see an ACF and PACF that each have a very modest first order spike. This could be interpreted as either an MA(1) or an AR(1). Try both. For part (f), the “comment” should be a look at diagnostic statistics (from sarima, for example) with a statement about the suitability of the model.

a. **There are slow increments (decrements) leading to sharp changes over a short time span along with an overall upward trend through the majority of the observed series. No other important features are of concern in the plot.**



b. **The ACF tapers, but levels off well above 0.**



c**. > fracdiff(x = x, nar = 1, nma = 0, M = 30)**

**Coefficients:**

**d ar ma**

**0.2525709 0.8758341 0.0000000**

**> fracdiff(x,nar=1,nma=0,M=30)$stderror.dpq**

**[1] 0.009824159 0.016153971**

**>temp=cbind(fracdiff(x,nar=1,nma=0,M=30)$d,fracdiff(x,nar=1,nma=0,M=30)$ar)**

**> temp/fracdiff(x,nar=1,nma=0,M=30)$stderror.dpq**

**[,1] [,2]**

**[1,] 25.70916 54.21788**

**d is significant (z=25.71 > 1.96)**

**phi\_1 is significant (z=54.22>1.96)**

d. **When the ACF of the time series does not taper to 0 suggesting a non-stationary series, we have found that taking first differences creates a stationary series.**

e. **Given the significant spike at lag1 followed by insignificant spikes, either an AR(1) or an MA(1) seem appropriate for the differenced data.** 

f. **ARIMA(1,1,0)**

**Coefficients:**

**ar1 constant**

**0.1698 -0.0016**

**s.e. 0.0312 0.0381**

**sigma^2 estimated as 0.9994: log likelihood = -1417.24, aic = 2840.49**

**$AIC $AICc $BIC**

**[1] 1.003419 [1] 1.005444 [1] 0.01323499**



**ARIMA(0,1,1)**

**Coefficients:**

**ma1 constant**

**0.1742 0.0059**

**s.e. 0.0314 0.0371**

**sigma^2 estimated as 0.9989: log likelihood = -1416.96, aic = 2839.93**

**$AIC**

**[1] 1.002858**

**$AICc**

**[1] 1.004883**

**$BIC**

**[1] 0.01267393**



**Both models result in significant Ljung-Box statistics and increased residual autocorrelation around lag 14. Because the ACF is just barely significant and this is the only issue at such a high lag, it is reasonable to assume this is just noise and that the model for the differenced data is suitable. When we extend this plot for higher lags, we see this is the only point of issue.**

**It is important to note that first differences performed well for a simulated ARFIMA model. Often both will perform well and you must decide between them. If the lag 1 ACF is near 1 and decays very slowly, this suggests that first differences are necessary. If the ACF decays very slowly for a lag 1 ACF less than 1, then an ARFIMA model may be useful.**

View question 2 below as a continuation of problem 1.

2. For the data in problem 1 (problem 5.1 of the book), use a fractionally differenced model with no ar term (as in Lesson 13.1). The format of the command would be fracdiff(y,nar=0,nma=0,M=30). Determine the residuals and then create the ACF of the residuals. As the answer, give the ACF and PACF and write a brief interpretation.

For programming, use the code in Lesson 13.1.

Teaching note: We were told that this was fractionally differenced ARIMA(1,1,0) so after a model with just the fractional difference we might then expect to see AR(1) residuals.



**The ACF and PACF clearly suggest an AR(1).**

3. For this problem, use the dataset globaltmp2.dat from the Week 13 folder. Download it to the directory you’ve set up for course datasets. The data are annual global temperature deviations from 1880 to 2004.

a. Do a time series plot of the data. Give the plot.



b. Determine the ACF and PACF of the data and write a brief interpretation.

**The ACF is slowly tapering to zero and the PACF suggests an AR(1).**



c. Estimate a fractionally differenced model with no AR or MA terms (just the fractional difference). What is the estimated value of d (the difference parameter)?

**Call:**

**fracdiff(x = x, nar = 0, nma = 0, M = 30)**

**Coefficients:**

**d**

**0.4945899**

d. Determine the residuals. Then get the ACF and PACF of the residuals. Give those plots and write a brief interpretation. Specifically, does the model in part c seem to be suitable?

**This model provides a good fit to the data given that there aren’t any significant autocorrelations. The two barely significant correlations appear to be noise given that lower lags are not significant and that they are just barely significant.**



4. Do problem 5.10 on page 316 of the textbook. The data are in the file sunspots.dat in the Week 13 folder. The problem in the book is unstructured, so we’ll make it structured. Do the following -

**A look at the plot of the series shows that the high peaks are sharper than the low peaks. A threshold model seems appropriate as the rate of increase/decrease changes as the series oscillates.**



a. Use a threshold value of c = 65 and an AR(4) model in each of the two regions.

##Regression for values below the threshold

**Call:**

**lm(formula = x1 ~ P1[, 1] + P1[, 2] + P1[, 3] + P1[, 4])**

**Residuals:**

**Min 1Q Median 3Q Max**

**-16.3740 -3.8615 -0.8408 3.8728 25.3136**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.94410 0.69562 7.107 7.86e-12 \*\*\***

**P1[, 1] 1.89372 0.05898 32.108 < 2e-16 \*\*\***

**P1[, 2] -1.26002 0.11764 -10.711 < 2e-16 \*\*\***

**P1[, 3] 0.53712 0.11653 4.609 5.86e-06 \*\*\***

**P1[, 4] -0.26688 0.05880 -4.539 8.05e-06 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 6.168 on 317 degrees of freedom**

**Multiple R-squared: 0.915, Adjusted R-squared: 0.9139**

**F-statistic: 852.6 on 4 and 317 DF, p-value: < 2.2e-16**

##Regression for values above the threshold

**Call:**

**lm(formula = x2 ~ P2[, 1] + P2[, 2] + P2[, 3] + P2[, 4])**

**Residuals:**

**Min 1Q Median 3Q Max**

**-21.936 -5.460 -1.267 4.485 27.656**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 15.02553 3.17799 4.728 5.89e-06 \*\*\***

**P2[, 1] 1.58245 0.08088 19.565 < 2e-16 \*\*\***

**P2[, 2] -1.06810 0.14951 -7.144 6.10e-11 \*\*\***

**P2[, 3] 0.65518 0.15061 4.350 2.75e-05 \*\*\***

**P2[, 4] -0.37317 0.07613 -4.902 2.82e-06 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 9.477 on 128 degrees of freedom**

**Multiple R-squared: 0.9146, Adjusted R-squared: 0.912**

**F-statistic: 342.8 on 4 and 128 DF, p-value: < 2.2e-16**

****

****

b. Use a threshold value of c = 100 and an AR(4) model in each of the two regions.

##Regression for values below the threshold

**Call:**

**lm(formula = x1 ~ P1[, 1] + P1[, 2] + P1[, 3] + P1[, 4])**

**Residuals:**

**Min 1Q Median 3Q Max**

**-24.8304 -4.1717 -0.9559 3.9793 29.8099**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.58915 0.67292 6.820 3.44e-11 \*\*\***

**P1[, 1] 1.86373 0.04933 37.779 < 2e-16 \*\*\***

**P1[, 2] -1.23749 0.09845 -12.569 < 2e-16 \*\*\***

**P1[, 3] 0.61762 0.10055 6.143 1.99e-09 \*\*\***

**P1[, 4] -0.31797 0.05182 -6.136 2.07e-09 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 7.042 on 394 degrees of freedom**

**Multiple R-squared: 0.9389, Adjusted R-squared: 0.9383**

**F-statistic: 1513 on 4 and 394 DF, p-value: < 2.2e-16**

##Regression for values above the threshold

**Call:**

**lm(formula = x2 ~ P2[, 1] + P2[, 2] + P2[, 3] + P2[, 4])**

**Residuals:**

**Min 1Q Median 3Q Max**

**-18.86829 -5.39702 0.07364 4.39652 17.91755**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 4.4760 6.7746 0.661 0.511775**

**P2[, 1] 1.5455 0.1258 12.280 < 2e-16 \*\*\***

**P2[, 2] -0.8479 0.2252 -3.765 0.000431 \*\*\***

**P2[, 3] 0.4791 0.2159 2.219 0.030931 \***

**P2[, 4] -0.3000 0.1031 -2.908 0.005367 \*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 8.915 on 51 degrees of freedom**

**Multiple R-squared: 0.904, Adjusted R-squared: 0.8965**

**F-statistic: 120.1 on 4 and 51 DF, p-value: < 2.2e-16**





For each “attempt” give the regression output, the ACF and PACF of the residuals and a plot showing the actual and predicted values. Comment on the suitability of each model.

**Neither ACF plot of the residuals looks suitable, which is why an AR(5) may be more appropriate. Both plots of the predicted values look good except for slight overestimations of the high peaks.**

Don’t difference the data. Use Lesson 13.2 and the code in those notes for guidance.

c. Comment on which of the two threshold values seems to be better. (Compare the results of parts a and b).

**The ACFs of the residuals are similar. Whether or not the MSE is better depends on whether you are looking above or below the threshold. The R-squared is better for a threshold of 100 below the threshold and close for the model above the threshold. Neither is clearly better than the other; based on the R-squared, a threshold of 100 may be prefereable.**

Teaching comment: An AR(5) may be better than an AR(4) but we’ll take it easy on you so that you can (almost) use the code in Lesson 13.2 as is.