ANA 535 Forecasting

Final Examination Practice Questions

Dr. Marvine Hamner

Foundations of Forecasting

The final exam will have you considering, analyzing, and forecasting Australian beer production. Load and consider the data in the "beerProduction.xlsx" dataset. That dataset is also available as part of Aus_Production in the fpp3 package. You will follow the process for time series analysis and forecasting that we have put together over the term. We have followed that process for the Amtrak data in the Course Project.

As usual, we'll start by conducting an EDA. First, generate a time plot of beer production in Australia. Use the plot type as "l" for a line plot so you can see what is going on more easily. Keep in mind that we are talking about time series data now. That is, the data may have periodicity or other patterns that change answers relative to what we might consider for panel data. That's ok. Time series data have additional features or attributes that panel data doesn't have.

1. How many observations are there in the dataset? 218
2. How many variables are there in the dataset? 2
3. Are the quarterly dates initially in a "date" format? No
4. Are there any missing values in the dataset? No

Make an initial time plot of the data to see what it looks like.

5. From the time plots of the data there appear to be distinct time ranges in the data. The first time range is from inception to about 1974 when production is increasing every year. The second time range occurs when production appears to more or less level off, from 1974 to 2010. Does it make sense to take the mean of production during the years from inception to 1974? No
6. Consider some descriptive statistics for the beer data. For example, what is the mean production volume of beer (in millions of liters) during the years when production is more or less consistent? Answer the question with a "whole number," i.e. the number of whole mega liters. (mega means millions.). 427.6905
7. Plot the data. What pattern do these data **appear** to exhibit? (Hint, you may have to zoom into a few years, e.g. 4-5 years, to answer this question!)
    a. monotonically increasing
    b. monotonically decreasing
    c. quadratic or square-root x
    d. cubic
    e. exponential

f. annually periodic
g. quarterly periodic x
h. a more complex combination of patterns

Split the beer dataset into two parts and plot together.

8. Do the data in the first part of the time range, from inception to about 1974, **appear** to have any trend or seasonality? Select the best choice below.
    a. trend
    b. seasonality
    c. both trend and seasonality x
    d. neither trend or seasonality
9. Do the data in the second part of the time range, after about 1974, **appear** to have any trend or seasonality? Select the best choice below.
    a. trend
    b. seasonality
    c. both trend and seasonality x
    d. neither trend or seasonality
10. Consider a histogram of the beer data. Does the histogram look normal or nearly normal (given that it is real world data)? Nearly normal
11. Consider a boxplot of the beer data. Do there appear to be "outliers" in the data? Yes
12. Given what you know about these data, e.g. that there is periodicity in the data, are there really "outliers" in the data? No

Because of the disparity between the two time ranges, at one point below we'll consider each separately. When appropriate, I'll refer to the first time range as firstBeer and the second time range as secondBeer. Eventually, we'll return to analyzing the data as one time series. You'll want to pay attention to when these switches occur.

13. What does a lag plot show?
    a. a scatter plot of the data, y_t
    b. a scatter plot of the data with a given lag versus the current data, $y_{t-1}\ vs\ y_t$ x
    c. a time plot of the data, $y_t\ as\ a\ function\ of\ t$
    d. none of these choices
14. Consider a "lags" plot of the data for both a short period of time, e.g. from 2000 to 2010, and for the entire dataset. Are there lags that have a high degree of positive correlation?
15. If you answered "Yes" to the last question, which lags show a high degree of positive correlation?
16. What feature or attribute in the data are the positively correlated lags associated with?
    a. peaks in periodicity
    b. troughs in periodicity
    c. unit roots
    d. no particular feature or attribute

17. Plot preliminary ACF and PACF plots. Are these plots consistent with what you observe in the dataset?
18. What is represented in an ACF plot?
    a. Spikes at autocorrelated lags, i.e. spikes with the magnitude of the autocorrelation coefficient at lags $y_t \; and \; y_{t-1}$
    b. Spikes at inversely correlated lags in the data
    c. Spikes at interesting points of time in the data
    d. None of these choices
19. What is represented in a PACF plot?
    a. Spikes at inversely correlated lags in the data
    b. Spikes at interesting points of time in the data
    c. Spikes at partial autocorrelated lags, i.e. spikes that are correlated at $y_t \; and \; y_{t-k}$ after removing the effects for lags of $1, 2, 3, \ldots, k-1$
    d. none of these choices
20. What do you see in the preliminary plots of ACF and PACF? (Remember to keep it simple!)
    a. trend in the data
    b. periodicity in the data
    c. potential for an AR(p) model fit
    d. potential for an MA(q) model fit
21. What are two of the common unit root tests?
    a. the ACF plot and the PACF plot
    b. the histogram and the boxplot
    c. the ADF test and the KPSS test
    d. none of these choices
22. What is the difference between unit root tests and the ACF/PACF?
    a. Unit root tests determine if a unit root is present in a times series whereas the ACF/PACF consider lags in evaluating the potential for AR(p) and MA(q) models.
    b. Unit root tests evaluate whether the data are stationary or non-stationary whereas the ACF/PACF are used to plot lags and partial lags in the data.
    c. Both a and b
    d. None of these choices
23. Conduct preliminary unit root tests (hint ADF and KPSS tests). Do these tests indicate that _____?
    a. the data are stationary
    b. the data are non-stationary
    c. the tests are inconsistent
    d. none of these choices
24. Take the log() of beer and plot. Outside the differences in scale, does the plot of log(beer) look appreciably different than the time plot of the data? (Yes/No)
25. Considering the decomposition of the entire beer dataset, what does the decomposition reveal?
    a. trend
    b. seasonality

c. trend and seasonality
d. none of these choices

26. What tool would be best to determine what the best period is to use for validation / forecasting?
    a. Autocorrelation Function
    b. Augmented Dickey Fuller test
    c. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test
    d. Histogram
    e. Probability Distribution
    f. Plot of the time series
27. Not including the spike at lag-0, what is the largest spike, i.e. the most dominant periodicity?
    a. 1 (quarterly)
    b. 2 (semi-annually, every 6 months)
    c. 3
    d. 4 (annually)
    e. 5
28. We know (by now) that we'll need to do something to make the data stationary before we can build a reliable model for the beer data. But for now let's go ahead and construct a linear model for the beer data (be sure to include all the components you selected in questions before, e.g. trend). Use that model to answer the following questions. (Note, your answers to these questions offer a big clue about whether or not it is appropriate to actually build a linear model at this point, during the EDA, of a project.)
    a. Are all the coefficients in the model statistically significant? (Yes/No)
    b. Does the variation in coefficients for seasonality follow the pattern you expected based on the autocorrelation function? For example, is the period that had the biggest spike in the ACF also the biggest coefficient in your linear model? (Yes/No)
    c. Plot your model including a forecast for the validation period. Does your model **appear** to capture all the phenomena in the data? (Yes/No)
    d. What is the accuracy in terms of the root mean squared error you got with your linear model? Enter your answer as a whole number!

    Plot and consider your model's residuals. Use these plots to answer the following questions.

    e. Does the plot of the residuals versus time show homoscedasticity? (Yes/No)
    f. Are the residuals normally distributed? (Yes/No)

Let's start looking at the data using the two different time ranges.

Consider the data over two different time ranges.

29. Considering a linear model built for **the first time range**, has constraining the model to include only data in this time range result in a more accurate model? (Yes/No)

30. What is the MAPE and the RMSE for the constrained model, i.e. for the first time range? Enter your answer to two decimal places, e.g. 1.43 or 6.27.
31. Now do the residuals for the first time range look normal?
32. Does the plot of the "Innovation" residuals have any particular pattern?
33. Does the ACF plot show any significant spikes?
34. Considering a linear model built for **the second time range**, has constraining the model to include only data in this time range result in a more accurate model? (Yes/No)
35. What is the MAPE and the RMSE for the constrained model, i.e. for the second time range? Enter your answer to two decimal places, e.g. 1.43 or 6.27.
36. Now do the residuals for the first time range look normal?
37. Does the plot of the "Innovation" residuals have any particular pattern?
38. Does the ACF plot show any significant spikes?

Now, go back to using the entire dataset rather than two different time ranges. We don't consider using a validation period or test set very often to see how accurate the models we build are. Let's look at that first. We've done some differencing which causes NA's to appear in the data. First, replace those NA's with 0's so the code doesn't throw an error and crash. Once you've done the replace command you should make sure that it worked correctly too.

Let's set up a 3-year test set or validation period. 3-years times 4 quarters equals 12, so nValid = 12. The training set, or nTrain, will just be the entire dataset minus the test set. Use the values of nValid and nTrain to split the dataset as desired using the window() command. Start working with the training set, first by decomposing it. Then, take a trend and a seasonal difference and answer the following questions.

39. Re-plot the beer production time plot. Then, plot the adjusted beer production data as a time plot. Consider these plots. Does it look like taking a trend and a seasonal difference was effective in making the data stationary? (Hint, think about what data should look like in a time plot if it is stationary.) (Yes/No)
40. We're calling the trend and seasonally adjusted data object beer.adj. Assume that the data are stationary and build a linear model. Plot the residuals and the ACF for the adjusted data. What phenomenon, if any, is/are in the adjusted data?
    a. trend
    b. periodicity
    c. trend and periodicity
    d. none of these choices
41. Use the model to forecast a 3-year period that corresponds to the time of the validation period. Generate a plot of the validation period data and the validation period forecast. Zoom into consider the differences. What phenomenon, if any, is/are in the adjusted data?
    a. trend
    b. periodicity
    c. trend and periodicity
    d. none of these choices

42. Generate a plot of the residuals, the ACF, and the histogram of the residuals for the adjusted beer data. Remember that is the data are stationary the plot of residuals should show that the residuals are "iid," the ACF should not have significant spikes, and the histogram should be normally distributed. Which of the following are true of this plot using gg_tsresiduals? Select all that are true.
    a. plot of residuals is "iid"
    b. ACF has no significant spikes
    c. histogram shows residuals are normally distributed
    d. none of these choices
43. What is the p-value of the ADF test on this data?
44. How many differences does the unitroot_ndiffs() command indicate we need to take?
    a. 0
    b. 1
    c. 2
    d. 3
    e. 4
45. Go ahead and take another seasonal difference from the data. Build the linear model and plot the residuals. The residuals actually look _____ .
    a. better
    b. worse
    c. about the same
    d. none of these choices
46. Check both unit root tests to see if they are consistent with your impression of the residuals. The ADF test and the KPSS test _____ .
    a. are inconsistent with the ADF test indicating the data are stationary
    b. are inconsistent with the KPSS test indicating the data are stationary
    c. are consistent indicating that the data are stationary
    d. are consistent indicating that the data are non-stationary
47. Using the diff() command from the R base package, take two seasonal differences from the beer data. Build a linear model using the twice differenced data. Are any of the coefficients of this linear model significant? (Yes/No)
48. Plot the residuals from this model using the gg_tsresiduals() command. Compare the plots in this figure to the plots generated from the previous twice differenced data. How do these plots of the residuals compare?
    a. they are quite similar; plot of residuals show residuals are iid, histogram of residuals show they are (reasonably) normally distributed, the ACF has some small but significant spikes
    b. only the plot of residuals is similar showing the residuals are iid, histogram of residuals show they are somewhat normally distributed, the ACF has some small but significant spikes
    c. nothing about these plots is similar, the plot of residuals, the histogram of residuals and the ACF are all quite different from the previous plot of residuals for the twice differenced data
    d. none of these choices

49. Although the unitroot_ndiffs() command indicates that we do not need to take any more differences, the PACF indicates _____ .
    a. a consistent result, the data are stationary
    b. spikes that all are not significant
    c. some significant spikes indicating that the data are still not quite stationary
    d. none of these choices
50. Both unit root tests, the ADF test and the KPSS test, indicate _____ .
    a. are inconsistent with the ADF test indicating the data are stationary
    b. are inconsistent with the KPSS test indicating the data are stationary
    c. are consistent indicating that the data are stationary
    d. are consistent indicating that the data are non-stationary
51. We have done everything that we are supposed to do to make the data stationary. We still have some indication that the data are not quite stationary but the spikes in the PACF are not too strong. Both unit root tests, the ADF test and the KPSS test indicate that the data are stationary. What does it mean that none of the coefficients in the linear model are statistically significant?
    a. the model is a good fit for the data
    b. the model is not a good fit for the data
    c. none of the predictor variables in the model are significantly related to the dependent variable (outcome)
    d. only a few of the predicator variables in the model are significantly related to the dependent variable (outcome)

We know this is a bad result. This would be the same bad result you might have seen in other courses for panel data. It is not bad because this is time series data. We have not built a model that really fits the data because none of the independent (predictor) variables are statistically significantly related to the dependent (outcome) variable. But, if we were paying attention we should have seen this coming. Reconsider the plot of the forecast for the validation period after we had removed the trend and the periodicity. Plotting the forecast against the original data shows that we have removed virtually all the characteristics present in the data. The model should not fit the data!!! This is an important result to think about. We did everything correctly but still came up with a very incorrect result!!! This is important to keep in mind as you progress through your career as an analyst.

What is the answer? Well, we'll move onto considering ARIMA models for which we do not have to remove the trend and periodicity in order to build a model. Let's see what we get from this. First, use best practices and tie this back to the beginning of our prior analysis. Generate time plots of the beer production data using both plot() and a time series data object and autoplot() and a tsibble. Make sure that you get the same plots you started with!

52. Now fit an ARIMA model to the data using the auto.arima() command in the forecast package. The criterion this program uses to select the best ARIMA model is _____ .
    a. AIC
    b. AICc
    c. BIC
    d. ADF
    e. ACF
    f. Other
53. From the model output what is the ARIMA model the auto.arima() command built?
    a. (1,1,1)(0,1,1)[4]
    b. (1,1,2)(0,0,1)[12]
    c. (1,1,2)(0,1,1)[4]
    d. (1,1,3)(0,1,2)[12]
54. This corresponds to an AR(1), MA(2), model with 1 (one) difference and a quarterly frequency.
    a. True
    b. False
55. The "best" ARIMA model selected by the program has _____. Select all that apply.
    a. trend
    b. seasonality
    c. trend and seasonality
    d. none of these choices
56. According to the output from your automatically generated ARIMA model, how many differences are taken?
    a. 1
    b. 2
    c. 3
    d. 4
57. Plot your forecast of the validation period (3 years). Considering your plot do you see that: (Select the best of the choices below.)
    a. the forecast values are less than the actual peaks and greater than the actual troughs
    b. match the actual data perfectly
    c. do not capture the trend correctly
    d. do not capture the seasonality correctly
58. When we stopped our analysis with the twice differenced data before we found that the resulting forecast no longer fit the data. Is this still true? (Yes/No)
59. Considering the output from your automatically generated ARIMA model, how many "seasons" are included in the model? (Note that the number of seasons has a relationship with the calendar year, e.g. if the data are monthly and the number of "seasons," or time periods, is reported as 12 then the seasonality is annual.)
    a. 1
    b. 2
    c. 3

d. 4

60. Based on the output accuracy for your auto ARIMA model, comparing it to the linear model you generated before, is the auto ARIMA model _____ ? (Hint: use the output from the model with 1 difference, the same as the ARIMA model. That model we named fit.beer.deseas.tslm.)
    a. the same level of accuracy as the linear model (or very close to it)
    b. better, or more accurate than the linear model
    c. worse, or less accurate than the linear model
    d. other
61. Consider a lags plot for the model you just built. Does this lags plot show the same result as before? That is, do the same lags show a strong positive correlation? (Yes/No)
62. Is the same periodicity shown in the ACF plot? (Yes/No)
63. Generate a 4-month moving average for the beer production data. Plot that moving average on a zoomed in portion of the data. What does the 4-month moving average illustrate?
    a. It follows the overall mean of the data.
    b. It illustrates times when both peaks and troughs are less, i.e. peaks are lower and troughs are higher.
    c. Both a and b.
    d. It has no bearing on the data so doesn't really illustrate anything.
64. Generate mean, naïve, and seasonal naïve forecasts for the 3-year validation period in the data, i.e. 2008 to 2010. What level does the mean forecast follow?
    a. the lowest point of the troughs
    b. the highest point of the peaks
    c. the mean or average of the last data point in the training period
    d. none of these choices
65. Generate mean, naïve, and seasonal naïve forecasts for the 3-year validation period in the data, i.e. 2008 to 2010. What level does the naïve forecast follow?
    a. the lowest point of the troughs
    b. the highest point of the peaks corresponding to the last data point prior to the validation period
    c. the mean or average of the last data point in the training period
    d. none of these choices
66. Generate mean, naïve, and seasonal naïve forecasts for the 3-year validation period in the data, i.e. 2008 to 2010. What level does the seasonal naïve forecast follow?
    a. the lowest point of the troughs
    b. the highest point of the peaks corresponding to the last data point prior to the validation period
    c. the mean or average of the last data point in the training period
    d. none of these choices
67. Add a naïve forecast with drift to your previous plot. Which forecast **appears** to best follow the data in the validation period?
    a. the mean forecast
    b. the naïve forecast

c. the naïve forecast with drift
d. the seasonal naïve forecast

68. What are the MAPE and RMSE values for the accuracy of the seasonal naive forecast of the validation period from this plot?
    a. 2.77 15.21
    b. 3.17 14,30
    c. 13.0 58.90
    d. none of these choices
69. Plot the residuals for the seasonal naïve forecast for the validation period. Does the histogram of the residuals look normal now? (Yes/No)
70. Does the plot of the residuals show that the residuals are "iid," independent and identically distributed? (Remember that if the residuals are independent there won't be any pattern in them. If the residuals are identically distributed they will be equally distributed about a mean equal to zero.) (Yes/No)
71. Are there any significant spikes in the ACF plot? (Yes/No)
72. As before, the lag = _____ is still a significant lag.
    a. 1
    b. 2
    c. 3
    d. 4
73. This would seem to indicate that there is still some seasonality in the data. However, the magnitude of the spike is relatively small. Check the unit root tests, i.e. the ADF and the KPSS tests. What do the results of these tests reveal?
    a. The tests are consistent that the data are stationary
    b. The tests are consistent that the data are non-stationary
    c. The tests are inconsistent. The ADF test indicates that the data are non-stationary
    d. None of these choices
74. Make a seasonal naïve forecast using recent_production) for 2-years in the future. Which of the following best describes this forecast (predicted) relative to the beer data (actual)?
    a. the forecast follows the data perfectly
    b. the forecast peaks appear higher and troughs lower than the data
    c. the forecast peaks appear lower and troughs higher than the data
    d. none of these choices
75. For the next model built, the tslm model, fit_beer, all the coefficients in the model are _____ .
    a. not significant
    b. significant
    c. significant only at the 0.1 level
    d. none of these choices
76. A plot of the fitted versus actual values for the tslm model, fit_beer, reveals that _____ .
    a. Fitted values in earlier years more closely follow the data.
    b. Fitted values in later years more closely follow the data.
    c. Fitted values do not follow the data at any time.

d. None of these choices

77. The plot of fitted versus actual values for Australian quarterly beer production by quarter indicates that most fitted values _____ .
    a. closely follow the actual data values
    b. do NOT follow the actual data values
    c. are inversely correlated with the actual data values
    d. none of these choices
78. The last two plots produced by the script shows the ets() and the seasonal naïve() forecasts for the validation period, and the stl() decomposition (season and trend decomposition using Loess) for the validation period. Consider the plot from 2000 through the end of the validation period 2010. Do the forecasts and stl() decomposition appear to follow the data fairly well? (Yes/No)
79. Zooming in on the last plot (from 2007 to 2010), the _____ appears to capture the peak values best.
    a. ets() forecast
    b. seasonal naïve forecast
    c. stl()
    d. none of these choices
80. Zooming in on the last plot (from 2007 to 2010), the _____ appears to capture the trough values best.
    a. ets() forecast
    b. seasonal naïve forecast
    c. stl()
    d. none of these choices